(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 July 2003 (17.07.2003)

PCT

(10) International Publication Number
WO 03/058868 A2

(51) International Patent Classification[7]: H04L

(21) International Application Number: PCT/US03/00163

(22) International Filing Date: 3 January 2003 (03.01.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/345,834 4 January 2002 (04.01.2002) US
60/384,438 31 May 2002 (31.05.2002) US

(71) Applicant *(for all designated States except US)*: **EINFINITUS TECHNOLOGIES, INC.** [IN/IN]; Harshada"A", 127/2, Mahaganesh Colony, Paud Road, Kothrud, Pune 411029, Maharashtra (IN).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **TANDON, Siddharth** [IN/US]; 870 East El Camino real, # 321, Sunnyvale, CA 94087 (US). **BANSAL, Jayant** [IN/IN]; Harshada"A", 127/2, Mahaganesh Colony, Paud Road, Kothrud, Pune 411029, Maharashtra (IN).

(74) Agent: **COLEMAN, Brian, R.**; Perkins Coie LLP, 101 Jefferson Drive, Menlo Park, CA 94025 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 03/058868 A2

(54) Title: DYNAMIC ROUTE SELECTION FOR LABEL SWITCHED PATHS IN COMMUNICATION NETWORKS

(57) Abstract: The invention teaches a method for dynamically routing data in a Multi Protocol Label Switching network. A network of nodes operative to receive and transmit data are grouped into a plurality of clusters which are hierarchically ranked; a leader node is also selected for each cluster. Data received at an ingress node is transmitted either to an egress node within the cluster, or, via the lead node, to another cluster of a different hierarchical rank. In some embodiments, a leader node contains in an associated routing table path information for each node in lower ranked clusters, as well as routing information for a leader node in at least one higher ranked cluster. The plurality of clusters may be organized hierarchically according to a rank determined by one or more Quality of Service (QoS) parameters, which may include jitter, loss, delay, and available bandwidth. The clusters may evolve along with traffic changes in the network; in particular, a given cluster may split or merge in response to changes in the applicable QoS metrics.

DYNAMIC ROUTE SELECTION FOR LABEL SWITCHED PATHS IN COMMUNICATION NETWORKS

BRIEF DESCRIPTION OF THE INVENTION

The present invention relates to the field of communications networks, and more particularly to systems and methods for providing dynamic route selection for label switching paths in a network.

BACKGROUND OF THE INVENTION

A Multi-Protocol Label Switching (MPLS) network architecture is a computer network architecture of connected nodes that conform to the official MPLS protocol, thereby enabling routing of data packets across such networks. The MPLS protocol architecture supports two options for route selection: (1) hop by hop routing, and (2) explicit routing. Hop by hop routing allows each node to independently choose the next hop for each data packet, and is the usual mode today in existing IP networks. A "hop by hop routed LSP" is a Label Switched Path (LSP) whose route is selected using hop by hop routing.

In an explicitly routed LSP, each LSR (Label Switched Router) does not independently choose the next hop; rather, a single LSR, generally the LSP ingress or the LSP egress, specifies several or all of the LSRs in the LSP. If a single LSR specifies the entire LSP, the LSP is "strictly" explicitly routed. If a single LSR specifies only some of the LSP, the LSP is "loosely" explicitly routed.

The sequence of LSRs followed by an explicitly routed LSP may be chosen by configuration, or may be selected dynamically by a single node. For example, the egress node may make use of the topological information learned from a link state database in order to compute the

entire path for the tree ending at that egress node. Explicit routing may be useful for a number of purposes, such as policy routing or traffic engineering. In MPLS, the explicit route needs to be specified at the time that labels are assigned, but the explicit route does not have to be specified for each IP packet. This makes MPLS explicit routing much more efficient than the alternative of IP source routing.

PRIOR ART FIG. 1 illustrates an explicit routing process used in existing MPLS networks. The process begins 52 with a data packet being received by an ingress node of the MPLS network. The data packet will include a header indicating the size and destination of the data packet. The ingress node will determine a path (LSP) to an intended egress node 54. The ingress node subsequently writes a label to the data packet 56 enabling the data packet to travel through subsequent hops. The ingress node then transmits the data packet to a next hop 58. If the next hop is not the egress node 60, the packet is transmitted to the next hop indicated by the LSP 62. A receiving node transmits the data packet to the next node of the LSP by adding a new label to the data packet header. If the node is the egress node 60, then the data packet is transmitted beyond the network by the egress node 64. A receiving node transmits the data packet to the next node of the LSP by adding a new label to the data packet header.

One problem with explicitly routing with externally configured paths is that changes in network topology (for example, extra traffic, additional number of users, changing QoS parameters, changes in link capacities, etc.) are not taken into account. This results in inadequate service levels. When explicit routing is used with externally configured paths to provide network services, such as VPNs, leased line etc., static paths are generated to satisfy the service requirements for each customer. These static paths remain underutilized, as customers who subscribe to services such as VPNs etc, do not use the connection 100% of the time. Additionally, customers tend to

overprovision bandwidth, so that the upgrades are required less frequently. The bandwidth wasted by use of static paths precludes the network from deploying a substantial number of customers, and diminishes the service provider's rate of return.

Additional problems with externally or manually configured explicit routing include the considerable time required to provision a new customer, or to upgrade/downgrade an existing customer. Moreover, in explicit routing with externally configured paths, 1:1 backup is generally maintained to provide for fault tolerance, which diminishes the network capacity available for use by 50%. Structured fault notification does not exist in existing MPLS networks, resulting in higher recovery times and delays on fault localization and identification. Existing provisions for simplifying fault localization and fault identification in MPLS network are inefficient and fail to meet the requirements of a commercial routing system.

SUMMARY OF THE INVENTION

The invention teaches systems and methods for dynamically routing data in an MPLS network, in a manner that responds to changes in network topology in real time. The invention can evolve in immediate response to changes in network characteristics, non-limiting examples of which include changes in traffic, increase/decrease in number of network users, evolving QoS parameters, or changes in link capacities. The dynamic routing enabled by the invention supports dramatic improvements in service levels and the number of users which may be supported on the network.

In embodiments of the invention, the communications nodes in a network, such an MPLS network, may be grouped into a plurality of clusters; the plurality of clusters may be ranked hierarchically, with a leader node is selected for each cluster. In some such embodiments, the leader node is selected on the basis of maximum connectivity to the other nodes of the cluster. The hierarchical rank may be based on measurements of one or more Quality of Service (QoS) parameters for the nodes, such that in a given cluster, each node meets a stated QoS threshold. As the QoS characteristics of the network evolve, the clusters may merge or split accordingly to remain in compliance with the applicable thresholds. The QoS parameters may include such network metrics as jitter, delay, loss, and bandwidth availability. In some embodiments, more than one QoS parameter is used to generate the cluster hierarchy. In some such embodiments, these QoS parameters are lexically ordered; in other embodiments, a weighted, normalized average of the QoS parameters may be taken.

In embodiments of the invention, data received at an ingress node is transmitted either to an egress node within the cluster, or, via the leader node, to another cluster of a different hierarchical rank. In some embodiments, leader nodes include routing tables populated with path information

for each node in lower ranked clusters, as well as path information for a leader node in at least one higher ranked cluster. These and other embodiments are described in greater detail herein.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a flow chart diagram of an explicit or manual routing method for use in MPLS networks in accordance with the prior art;

FIG. 2 is a block diagram of a communications node in accordance with an embodiment of the present invention;

FIG. 3 is a flow chart of a process for creating an MPLS network hierarchy in accordance with an embodiment of the present invention;

FIG. 4 is a flow chart illustrating a process for selecting a leader node for a cluster in accordance with an embodiment of the present invention;

FIG. 5 is a flow chart illustrating a process for creating routing tables for each node in the MPLS network in accordance with an embodiment of the present invention;

FIG. 6A is a schematic block diagram of constellations comprising a first level of the hierarchy of constellations in accordance with one embodiment of the present invention;

FIG. 6B is a schematic block diagram of constellations comprising a second level of the hierarchy of constellations in accordance with one embodiment of the present invention;

FIG. 6C is a schematic block diagram of constellations comprising a third level of the hierarchy of constellations in accordance with one embodiment of the present invention;

FIG. 6D is a schematic block diagram of constellations comprising a fourth level of the hierarchy of constellations in accordance with one embodiment of the present invention;

FIG. 6E is a schematic block diagram of constellations comprising a fifth level of the hierarchy of constellations in accordance with one embodiment of the present invention;

FIG. 7 is a flow chart diagram generally illustrating a process for routing data in accordance with one embodiment of the present invention;

FIG. 8 is a schematic block diagram generally illustrating a process for re-executing a process to form new clusters in accordance with one embodiment of the present invention;

FIG. 9A is a schematic block diagram generally illustrating a group of constellations about to undergo a split in accordance with one embodiment of the present invention;

FIG. 9B is a schematic block diagram generally illustrating a group of constellations of the second level about to undergo a split in accordance with one embodiment of the present invention;

FIG. 9C is a schematic block diagram generally illustrating a group of constellations of the third level about to undergo a split in accordance with one embodiment of the present invention;

FIG. 9D is a flow chart diagram generally illustrating a process for executing a split in response to changes in network topology in accordance with one embodiment of the present invention;

FIG. 9E is a schematic block diagram generally illustrating the group of constellations of FIG. 9A about to undergo a split in accordance with one embodiment of the present invention;

FIG. 9F is a schematic block diagram generally illustrating the nodes of FIG. 9E after a split is performed in accordance with one embodiment of the present invention;

FIG. 9G is a schematic block diagram of constellations comprising the second level of the hierarchy of constellations of FIG. 9F;

FIG. 9H is a schematic block diagram of constellations comprising a third level of the hierarchy of constellations of FIG. 9A;

FIG. 10A is a schematic block diagram generally illustrating an exemplary group of constellations about to undergo a merger in accordance with one embodiment of the present invention;

FIG. 10B is a flow chart diagram generally illustrating a process for executing a merger of constellations in response to changes in network topology in accordance with one embodiment of the present invention;

FIG. 10C is a schematic block diagram generally illustrating the result of a merger of the constellations illustrated in FIG. 10A;

FIG. 11 is a table diagram generally illustrating a routing table at in accordance with the present invention; and

FIG. 12 is a table diagram generally illustrating a Partial Forwarding Table in accordance with one embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

Overview

The present invention teaches a new paradigm for efficiently and dynamically routing data in a communications network. The invention enables the network to evolve dynamically in immediate response to changes in network performance, thereby improving link utilization, maximizing users, and improving support levels. The dynamic routing enabled by the invention supports dramatic improvements in service levels and the number of users which may be supported on the network.

FIG. 2 illustrates one possible representation of a communications node in a network in accordance with one embodiment of the present invention. Communication node 102 is coupled to communication links 104, 106, 108, 110. Each communications link 104, 106, 108, 110 is operative to carry a predetermined transmission bandwidth of data, voice or video information. Each communications link 104, 106, 108, 110 has a respective transmission bandwidth, latency, jitter and packet loss ratio associated with that link. Note that these parameters may be a function of time.

Generally node 102 may be either a redistribution point, or an end point (terminal node) for data transmissions. Terminal nodes, such as phone sets, computers, printer or fax machines, generate or use information transmitted over the network, or facilitate communications with other networks. Communications nodes may include networking equipment such as switches, routers, or gateways, which are operative to recognize and forward transmissions to other nodes. In accordance with one embodiment of the present invention, the node 102 has a software agent residing in the control plane of a router, which is responsible for calculating node hierarchy and routing tables.

Creation of Hierarchically-Ordered Node Clusters

FIG. 3 illustrates a process 150 for creating an MPLS network hierarchy in accordance with one embodiment of the present invention. The process 150 of creating a hierarchy of clusters of MPLS nodes is herein referred to as an X-Constellation algorithm, with the term "cluster" used interchangeably with the term "constellation" herein. The X-constellation algorithm process 150 begins 152 with all nodes meeting both a first primary quality of service (QOS) parameter threshold and any secondary QOS parameter thresholds grouped into clusters in the first tier. Clusters are composed of all nodes that are connected with each other via links meeting the first primary QOS and any required secondary QOSs thresholds. For example, all nodes connected contiguously via a 2Mb/s or greater bandwidth link could be grouped in a first cluster of the first tier, with no secondary QOS parameter requirements for the cluster. For example, a node having four links of 1Mb/s, 2Mb/s, 1Mb/s and 1Mb/s would be included in the first tier as having a 2Mb/s bandwidth and would be in a cluster with all other nodes with which it could maintain a 2Mb/s bandwidth path. In alternative embodiments primary or secondary QOS parameters may include bandwidth, latency, jitter, packet loss ratio, topological area, policy settings, etc.

In a subsequent step 154 all nodes in the network falling below the first primary QOS threshold and meeting a second primary quality of service (QOS) parameter threshold are grouped in clusters of a second tier. As an example, all nodes connected contiguously via links having between 1Mb/s and 2Mb/s bandwidth may be grouped in a cluster.

In the next step 156, all nodes falling below the second primary QOS threshold and meeting a third primary quality of service (QOS) parameter threshold are grouped into clusters of the third tier. In the present example, all nodes contiguously connected via links having between 0Mb/s and

1Mb/s bandwidth are grouped in a cluster. In this non-limiting example, all remaining nodes falling into a fourth tier.

In a subsequent step 160, all of the clusters are arranged in an ascending hierarchy according to tier rank. In the current example, the fourth tier is above the third tier and so on down to the first tier. After arranging the clusters hierarchically, a leader node is selected for each cluster 162. In embodiments of the invention, leader nodes generally have greater communication capacity than the non-leader nodes within a particular cluster; this is described in greater detail infra.

Populating the Routing Tables

Upon selection of the leader nodes, routing tables are calculated for each node in the network 162. In embodiments of the invention, the algorithm for populating the table may include:

- Calculating routing tables within the constellations of the actual network
- Calculating routing tables for the virtual network, i.e., for all levels above actual network level.

Once the network has been mapped hierarchically according to QoS parameters as described above, the routing algorithm is called by each node in network which calculates the routing table used by LDP and subsequently by MPLS. In embodiments of the invention, the routing algorithm may also include the following:

For each node

- Make adjacencies to each other node in the constellation
- Calculate the shortest path to each other node in the constellation

- Calculate the minimum bandwidth link to reach all other nodes in the constellation, and store this bandwidth information in the routing table with the respective destination entry
- Determine the leader in the constellation, and mark a leader bit to the destination entry in the routing table

Note that the shortest path may be calculated using Dijkstra's algorithm, or any other shortest path algorithm known to those skilled in the art.

In embodiments of the invention, records in a routing table generally includes: destination address of the packet; next hop; bandwidth available to next hop; hierarchy of next hop; leader bit of next hop; and an adjacency bit. For non-leader nodes, the routing table includes the shortest path to each other node in the cluster and to every adjacent node regardless of cluster. In embodiments of the invention, the leader node's routing table also includes the shortest path to each node of all lower clusters, and the shortest path to the leader node of the next higher cluster. For example a non-leader node of the third cluster will have a routing table including the shortest path to each node of the third cluster including the leader node of the third cluster. A leader node's routing table will include the shortest path every node of the leader node's cluster and the shortest path to each node belonging to lower clusters, and the shortest path tc the leader node of the next higher cluster.

Though the above examples deal with quality of service QOS thresholds based solely on bandwidth requirements, it should be understood that virtually any other parameter may be used with the X-constellation algorithm to generate node cluster hierarchies. Examples of quality of service parameters that may be used to create cluster hierarchies include but are not limited to,

bandwidth, latency, jitter, packet loss ratio, topological area, and policy settings. Other suitable QOS metrics shall be apparent to those skilled in the art.

Additionally more than one parameter may be used to generate X-Constellation hierarchies. In an embodiment of the invention, the primary quality of service parameter may be the cross-product of one or more of the node parameters. For example, the primary QOS parameter used in steps 152 through step 158 may be the product of bandwidth, latency and jitter. Alternatively both the primary and secondary quality of service parameters could be cross products of bandwidth, latency, jitter, packet loss ratio, topological area, policy settings, etc. In some embodiments of the invention, a lexical ordering of one or more of the QoS parameters may be used to generate the hierarchy. In alternative embodiments, a normalized weighted average of one or more of the QoS parameters may be used to generate the hierarchy. Other alternative formula combining any basic parameters will be apparent to those skilled in the art, and any combination thereof may be used as a primary or secondary quality of service parameter in steps 152 through 158 of the X-Constellation algorithm.

FIG. 4 illustrates a process 162 for selecting a leader node for a cluster in accordance with another embodiment of the present invention. The process begins 164 by determining which nodes in a cluster have the most links. If more than one node in the cluster has the highest number of communications links 166, the bandwidths of all the communications links is summed for each such node. The candidate node with the greatest total bandwidth of communications links is then selected as the leader node for that cluster. 170. In embodiments of the invention, the leader bit of the selected leader node is assigned a value of 1, signifying that the node is a leader node172.

If only one node of the cluster has the highest number of communications links 166 the node is selected as the leader node for that cluster 174. The leader bit of the selected node is assigned a

value of 1 to signify that the node is the leader node for the cluster 176. All other nodes in the cluster will have a leader bit value of 0 to signify that they are not leader nodes.

FIG. 5 illustrates a process 163 for creating routing tables for each node in the MPLS network in accordance with one embodiment of the present invention. Initially, the leader bit of a node is read 502. If the node is not a leader, a routing table is calculated for the node 504. The routing table for a non-leader node will contain the shortest path from the non-leader node to each other network node that is either one hop from the current node or in the same cluster as the current node. For example, a non-leader node of a cluster of nodes will have a routing table including the shortest paths to any adjacent node and to each node of that cluster, including the leader node of that cluster. Routing table entries may further include a destination address, next hop, bandwidth available to next hop, hierarchy of next hop, leader bit of next hop, and adjacency bit for each of these paths.

In embodiments of the invention, a Partial Forwarding Table (PFT) and a Label Information Base (LIB) may be created for each path entry included with the node's routing table. The PFT may include records of an IP address of a router, a label information base pointer, a leader bit, a bandwidth parameter, and an alternative path for each path entry. Records in the LIB may include fields such as an index, incoming interface, incoming label, outgoing interface, outgoing label, and an LIB pointer. Many methods of populating the PFT and LIB will be apparent to those skilled in the art.

If a node is determined to be a leader node 502, a routing table is calculated for the leader node 512. This routing table may include the shortest path to each node that is one hop from the leader node, each node that is in the leader node's cluster, each node that is in a lower cluster, and the shortest path to the leader of the next higher cluster. For example, the leader of a cluster of

nodes may have a routing table including the shortest paths to any adjacent node, each node of the second cluster, each node of the first cluster, and the leader node of the third cluster.

In embodiments of the invention, a PFT and an LIB is created for each path entry included in the leader node's routing table 514. Records in the PFT may include one or more of the following: a destination address, next hop, bandwidth available to next hop, hierarchy of next hop, a leader bit of next hop, and an adjacency bit for each of these paths. The PFT of a node may also contain records including an IP address of a router, a label information base pointer, a leader bit, bandwidth, and alternative path. The LIB may include records containing an index, an incoming interface, an incoming label, an outgoing interface, an outgoing label, and a LIB pointer. Many methods of populating the PFT and LIB would be apparent to those skilled in the art.

Examples of Node Clusters

FIG. 6A-6E are schematic block diagrams illustrating examples of a hierarchy of constellations in accordance with one embodiment of the present invention. FIG. 6A is a schematic block diagram of constellations 600 comprising a first level of the hierarchy of constellations. In the non-limiting example shown herein, each unspecified link has a bandwidth capacity of at least 100Mbs. Each large oval represents the border of a distinct constellation (cluster). All nodes are contained within such a border because all nodes within the cluster have at least 100Mbs with one another. A first constellation 602 includes nodes 1, 2, 3 and 4 connected contiguously via links of at least 100Mbs. The leader node of constellation 602 is node 1. As shown in FIG. 6A, nodes 5 and 6 are not included in constellation 602, because the link connecting them to first constellation 602 has less than 100Mbs of bandwidth. As such, links 5 and 6 form a first level constellation 604, with node 5 as the leader of constellation 604. Constellations 606 and 608 are comprised of nodes 9, 13, 14 and 10, 15 respectively.

FIG. 6B is a schematic block diagram of constellations at 610 comprising a second level of the hierarchy of constellations. In the non-limiting example depicted herein, second level constellations are comprised of nodes with links having a capacity of less than 100Mbs and at least 75Mbs. A first constellation 612 of the second level includes nodes 1, 5 and 10, with node 1 being the leader of the constellation. A second constellation 614 of the second level is made up of nodes 9 and 12. As is shown in FIG. 6B, nodes 7 and 8 are not members of constellations at the second level of the hierarchy of constellations because they have no links of between 100Mbs and 75Mbs, though they are members of constellations at level one (FIG. 6A). As such, nodes 7 and 8 are represented as independent nodes at the second level.

FIG. 6C is a schematic block diagram of constellations at 620 comprising a third level of the hierarchy of constellations. In the present example third level constellations include all nodes with links having a capacity of less than 75Mbs and at least 50Mbs. A first constellation 622 of the third level includes nodes 1, 7 and 9, with node 1 being the leader of the constellation. In this example, there is only one constellation in the third tier.

FIG. 6D is a schematic block diagram of constellations at 630 comprising a fourth level of the hierarchy of constellations. In this example, fourth level constellations include all nodes with links having a capacity of less than 50Mbs and at least 25Mbs. The only constellation 632 of the fourth level includes nodes 1 and 8, with node 1 being the leader of the constellation.

FIG. 6E is a schematic block diagram of constellations at 640 comprising a fifth level of the hierarchy of constellations. In the present example, fifth level constellations contain all nodes with links having a capacity of less than 25Mbs. The only constellation 642 of the fifth level includes nodes 1 and 11, with node 1 being the leader of the constellation. As can be seen from FIG. 6A-6E,

a node may be a member of different clusters at different levels, and may be a leader node at multiple levels of the hierarchy.

Note that though the example provided above utilizes a five level hierarchy, any number of levels may be employed.

Routing Data Through the Node Clusters

FIG. 7 illustrates a process for routing data 700 in accordance with one embodiment of the present invention. Data is received by an ingress node of the MPLS network 702. An ingress node is a node communicatively coupled to systems outside the MPLS network. The data packet includes a header containing information indicating the data packet's destination. The ingress node will determine the data packet's egress node based on the packet's destination. The ingress node will compare the egress node with the ingress node's routing tables (which was generated in the process of FIG. 3). If the intended egress node for the data packet is in the same cluster as the ingress node (has an entry in the node's routing tables (see FIG. 3) for a non-leader ingress node) in step 704 the process continues to step 706. At 706 the ingress node transmits the data packet to the egress node by a method of label switching. The new label being created from the LIB created in the process of FIG. 5.

If the packet is intended to leave the MPLS network, the egress node transmits the packet to a destination outside the network in step 708, or the egress node may use the data packet in some manner.

If the ingress node determines that the egress node is not within the ingress node's cluster 704 (for e.g., if the ingress node is not a leader node, and there is entry in its routing table for the

egress node), the ingress node transmits the data packet to the leader node of the ingress node's cluster 710.

If the egress node is in a cluster below the cluster of the leader node 712, the leader node makes this determination by consulting its routing table 714. The leader node subsequently transmits the packet to the egress node. As discussed with reference to FIG. 3, the leader node has routing table entries for every node of every lower cluster. In step 716 the egress node uses the data packet or forwards it to a final destination beyond the MPLS network.

If the egress node is not below the cluster of the leader node 712, the QOS demand for the packet is compared to the QOS values of the next higher level 717. If the quality of service requirement for the packet is met by the next higher level of the MPLS network, the leader node transmits the data packet to the leader node of the next higher cluster 718. If the lead node is also the lead node of the above cluster, it simply keeps the packet.

The leader node which receives the packet determines whether the egress node is in the receiving leader node's cluster or a lower cluster 720. If the egress node is in the current cluster or a lower cluster, the leader node will transmit the data packet to the egress node 724. The egress node either utilizes the data packet, or proceeds to forward it out of the network 726.

If the egress node is not in the receiving leader node's cluster or a lower cluster 720, the process determines whether a higher cluster exists 722. If a higher cluster exists, the leader node transmits the data packet to the leader node of the next higher cluster 718.

If no higher clusters exist 722, the data packet is dropped 728. This occurs because there is no viable path to the egress node. In embodiments of the invention, a detailed error message to be returned in response.

Thus information packets will continually travel to higher and higher clusters until they reach a leader node that knows a path to the egress nodes. This greatly increases the efficiency of routing in MPLS networks by reducing the information management required by any single node when compared to a purely hop by hop routing method.

For example, a data packet entering the system at node 2 (FIG. 6A) with a final destination of node 6 (or destination at some point outside the MPLS network via node 6), will be transmitted to node 1 (the lead node of the first cluster 602) 710. Since node 6 is not in a cluster below node 1, the QOS of the second level of the network is compared to the packet's QOS requirements 717.

If the QOS requirement is met, the packet is sent to the lead node of the next higher level above the current lead node 718. In the present example, since node 1 is the leader above node 1 at the second level, the packet remains at node 1, but node 1 is considered to be at the second level. Node 1 checks its routing tables to determine if node 6 is below it 720. Since node 6 is below node 1 at the second level, node 1 sends the packet to node 6. The routing tables of node 1 will contain a shortest path to node 6, since node 6 is on a lower level than lead node 1 at the second level.

Evolution of Clusters

FIG. 8 illustrates a process at 800 for re-executing the X-constellation process of FIG. 3 in order to form new clusters in accordance with one embodiment of the present invention. The MPLS network initially executes the X-constellation algorithm 802 as illustrated in FIG. 3. A new routing table is populated according to the X-Constellation algorithm and the PFT and LIB are calculated on the basis of the routing table. Upon receipt of a data packet 804 the data packet travels through the MPLS network 808 via a path. The sum of the data flow along this path is then determined. If the sum of data flow, i.e., the bandwidth path across the MPLS network, exceeds a predetermined

threshold, the X-constellation grouping algorithm of FIG. 3 is re-executed 812, with the current capacities of nodes and links within the MPLS system used as input.

In embodiments of the invention, once a flow is allocated on a particular path, the dynamic bandwidth entry corresponding to each link of the path is updated to reflect the change in the bandwidth of that path, caused by allocating the flow. If the change in bandwidth exceeds the threshold value, then re-clustering is undertaken. Thus, with allocation and de-allocation of demands, existing clustering may split or merge.

FIGs 9A-9H illustrate an example of an MPLS network hierarchy undergoing a split in accordance with an embodiment of the present invention. FIG. 9A illustrates an initial group of constellations about to undergo a split 920. The nodes are connected via communications links; those links illustrated in FIG. 9A without a specified transmission capacity have at least 100 Mbs of transmission capacity. The first level of constellations include constellations 922, 924, 926, 928, 930, 932 and 934. In this example, a node 14 has been selected as the leader node of the first constellation 922.

Requests for bandwidth allocation may occur between two nodes either within a constellation or between two nodes of different constellations. Allocations occurring within a constellation may result in a split. An allocation of bandwidth on a link, wherein the remaining link bandwidth still meets the level bandwidth, does not result in a merger. The bandwidth change is simply propagated to the nodes of the present constellation as well as in the tree of all leader nodes presiding over the link undergoing allocation. For example, link 936 is allocated 10Mbs of data transmission, leaving 110 Mbs of capacity. Since the capacity of link 936 still exceeds the 100 Mbs threshold of the first level, no split is required. The new link bandwidth is communicated to node 14

(the leader of the present constellation), the leader node 14, all nodes within the present constellation, and to any leader nodes in linked constellations which are higher in the hierarchy.

FIG. 9B is a schematic block diagram of constellations at 940 comprising a second level of the hierarchy of constellations of FIG. 9A. Second level constellations are comprised of nodes with links having a capacity of less than 100Mbs and at least 75Mbs. A first constellation 942 of the second level includes nodes 5, 10, 12 and 14, with node 14 being the leader of the constellation 942. As is shown in FIG. 9B, nodes 7 and 8 are not members of constellations at the second level of the hierarchy of constellations because they have no links of between 100Mbs and 75Mbs, though they are members of constellations at level one (FIG. 9A). Instead nodes 7 and 8 are represented as independent nodes at the second level.

FIG. 9C is a schematic block diagram of constellations at 946 comprising a third level of the hierarchy of constellations of FIG. 9A. Third level constellations include all nodes with links having a capacity of less than 75Mbs and at least 50Mbs. A first constellation 948 of the third level includes nodes 7 and 14, with node 14 being the leader of the constellation. Only a single constellation of the third level exists in this exemplary embodiment.

FIG. 9D illustrates a process 850 for re-executing the X-constellation process in order to split constellations in response to changes in network topology in accordance with one embodiment of the present invention. A transmission bandwidth 852 is allocated to a link 936 within the MPLS network. The remaining bandwidth capacity of the link 936 is compared to the bandwidth requirement of the link 854. If the resulting bandwidth capacity retained by the link is insufficient for the level the link is currently on, an alternative path is checked 856. If no alternative path of sufficient bandwidth is available, a split is executed, creating two new constellations in step 858 separated by the link 936. New leaders are then selected for each new constellation as described

with regard to FIG. 4. The routing tables of all nodes of both new constellations and the leaders above each constellation are modified 860 to reflect the change in the network.

If, the resulting bandwidth of link 936 still qualified for the link's current level within the hierarchy of levels 854, the routing tables of all nodes within the constellation to which the link 936 is contained would be modified 862 to reflect the reduced bandwidth capacity of link 962. The routing tables of the leader nodes above the constellation level would also be modified to reflect the reduced capacity of link 962. Similarly, if an alternate path exists 856, the routing tables are modified accordingly 862. Note that though the process of FIG. 9D refers solely bandwidth requirements, the quality of service requirements may include many other parameters or a combination of parameters, as has been discussed previously herein and will be apparent to those skilled in the art.

FIG. 9E illustrates the group of constellations of FIG. 9A about to undergo a split at 949 comprising a first level of an MPLS network hierarchy in accordance with one embodiment of the present invention. The nodes are connected via communications links. Links illustrated in FIG. 9A without a specified transmission capacity have at least 100 Mbs of transmission capacity. Nodes of a first constellation 922 are connected contiguously via links of at least 100 Mbs capacity. As illustrated, link 936 has a reduced transmission capacity of 80 Mbs. Since no alternate route of at least 100 Mbs exists between node 9 and node 1, a split is performed.

FIG. 9F illustrates the nodes of FIG. 9E after a split is performed in accordance with one embodiment of the present invention. Constellation 922 from FIG. 9E is partitioned into constellations 951 and 952. Constellation 951 includes nodes 9, 13 and 14. Constellation 952 includes nodes 1, 2, 3 and 4. New leaders are selected for constellations 951 and 952. Node 1 is selected for constellation 952 and node 14 is selected for constellation 951. The routing tables of

each member of constellations 951 and 952 are modified to reflect the topology of the new constellations.

FIG. 9G is a schematic block diagram of constellations 970 comprising the second level of the hierarchy of constellations of FIG. 9F. Second level constellations are comprised of nodes with links having a capacity of less than 100Mbs and at least 75Mbs. A first constellation 972 of the second level includes nodes 1, 5, 10, 12 and 14, with node 14 being the leader of the constellation 942. The routing tables of the leader node 14 of constellation 972 are modified to reflect the changes in network topology caused by the split of FIG. 9F.

On any level above the first level, the links depicted may represent a path comprising many individual links and nodes. For example, link 936A is in fact a virtual link comprised of multiple links with the stated minimum bandwidth capacity.

FIG. 9H is a schematic block diagram of constellations 946 comprising a third level of the hierarchy of constellations of FIG. 9A. Third level constellations in this example are comprised of nodes with links having a capacity of less than 75Mbs and at least 50Mbs. A first constellation 948 of the third level includes nodes 7, 9 and 14, with node 14 being the leader of the constellation 942.

Whenever there is a split or merge of clusters, entries are created or destroyed from the tables of a router. For instance, when the constellations merge, new paths are formed within the constellation by setting up new LSP's and increasing the size of the PFT and LIB tables of a router. When there is a split, LSP's are destroyed and entries are deleted from the routing tables. Thus, the routing tables keep changing with allocation and de-allocation of demands.

FIG. 10A illustrates an example group of constellations 1000 which are about to undergo a merger. Constellations 1002 and 1004 are connected via links 936, 1006 and 1008. Link 936 has an

initial bandwidth capacity of 80 Mbs. As an illustrative example, traffic originally assigned to link 936 is de-allocated resulting in link 936 having a bandwidth capacity of 120 Mbs. As bandwidth is generally de-allocated when transmissions of previously assigned data are completed, and the capacity of link 936 would exceed the threshold of the first level hierarchy (100Mbs), this bandwidth assignment would result in a merger between constellations 1002 and 1004

FIG. 10B illustrates a process 1050 for re-executing the X-constellation process in order to merge constellations in response to changes in network topology in accordance with one embodiment of the present invention. At the outset, the transmission bandwidth of link 936 is de-allocated 1052. In step 1054 the resulting bandwidth capacity of the link 936 is compared to the bandwidth requirement of the current level of the link. If the resulting bandwidth capacity retained by the link is sufficient for a lower level within the hierarchy of levels, a merger is executed 1056, combining the constellations at each end of link 936 in order to form a new constellation. A new leader is then selected for the new constellation 1058. The routing tables of all nodes of the new constellation and the leaders above the new constellation are then modified 1060 to reflect the change in the network.

If the resulting bandwidth of link 936 does not qualify for a lower level within the hierarchy of levels 1054, then the routing tables of all nodes within the constellation to which the link 936 is contained would be modified 1062 to reflect the reduced bandwidth capacity of link 962. The routing tables of the leader nodes above the constellation level would also be modified to reflect the increased capacity of link 936.

FIG. 10C illustrates the result of constellations 1002 and 1004 merging at 1090 to form constellation 1092 due to the deallocation of traffic originally assigned to link 936, leaving link 936 with a bandwidth capacity of 120 Mbs. The merger process is performed in much the same way as

the splitting process of FIG. 9. After the new constellation 1012 is formed, the routing tables of each node of the constellation 1092, and the leader nodes directly above constellation 1092 are modified to reflect changes in network topology.

Routing Tables Used in Embodiments of the Invention

FIG. 11 illustrates a routing table at 1100 in accordance with one embodiment of the present invention. As discussed with reference to FIG. 7, the routing table of a non-leader node will include a shortest path meeting all quality of service requirements to every member of the non-leader node's cluster, the leader of the cluster and all adjacent nodes. A leader node's routing table will also include the shortest path (that meets all quality of service requirements) to the next higher cluster's leader node and all nodes belonging to lower clusters. For example, in a 4 cluster hierarchy, the third cluster leader node will have paths to the leader node of the fourth cluster, and all nodes of the first, second and third clusters.

FIG. 11 illustrates a Partial Forwarding Table at 1100 in accordance with the present invention. The PFT 1100 is used to determine a path for data packets initially entering the MPLS network. The FEC is the forwarding equivalence class (format of QOS requirement), usually bandwidth. PHB is per-hop forwarding behaviors (used in diffserv) not used in a preferred embodiment. LIBpter refers to LIB table entry to a corresponding FEC providing outgoing label and interface. Alternative Path is a pointer to the LIB table entry other than that of the LIBptr.

FIG. 12 illustrates a Label Information Base at 1200 in accordance with one embodiment of the present invention. The LIB 1200 is used to generate a label for each hop a data packet will take within the MPLS network. The entry iIface refers to an incoming interface, iLabel to incoming

label, oIface to outgoing interface, oLabel to outgoing label, and LIBpter is used to make a stack of labels.

The PFT table is used at the Ingress Node to forward a packet along established LSP. As the packet (unlabeled packet) arrives at the Ingress Node, FEC corresponding to the destination IP address is looked in the PFT table. If a matching FEC is found, the LIBptr of matching is used to forward packet on ongoing interface by applying the outgoing label, both pointed by LIBptr (which points to entry in LIB table). If entry is not there then the packet is forwarded using the entry having leader bit one.

At LSR as labeled packet arrives at the node, that label is looked in the LIB table. If a matching entry is found then the outgoing interface and outgoing label is determined using the matching entry. Outgoing label is attached to the packet and it is forwarded through the outgoing interface. If outgoing label is Zero then this node is Egress for this LSP the label is popped and forwarded to IP layer.

Alternative Embodiments and Conclusion

The present invention may be combined with, or applied to various other innovations relating to MPLS routing including: U.S. Patent No. 6,295,296, entitled, Use of a single data structure for label forwarding and imposition; U.S. Patent No. 6,275,493, entitled, Method and apparatus for caching switched virtual circuits in an ATM network; U.S. Patent No. 6,272,131, entitled, Integrated data packet network using a common time reference; and U.S. Patent No. 6,205,488, entitled, Internet protocol virtual private network realization using multi-protocol label switching tunnels, each of which is hereby incorporated herein by reference.

The present invention may also be applied to networks other than MPLS networks, as it is applicable to any network routing system capable of transmitting data through a plurality of links

and nodes. By way of non-limiting example, the invention may be applied to networks using many different data link media, such as ATM and Gigabit Ethernet. Routing protocols that may be used in conjunction with the present invention include BGP, OSPF, and RIP. Transport protocols that may be used with the invention include TCP/IP. The foregoing examples illustrate certain exemplary embodiments of the invention from which other embodiments, variations, and modifications will be apparent to those skilled in the art. The invention should therefore not be limited to the particular embodiments discussed above, but rather is defined by the following claims.

What is claimed is:

CLAIMS

1. A method for dynamically routing data in a Multi Protocol Label Switching (MPLS) network, the network including a plurality of nodes, each node operative to receive and transmit data, the method comprising:

grouping the nodes into a first plurality of clusters, wherein each cluster has a hierarchical rank based on a measurement of one or more QoS metrics for each of the first plurality of clusters;

selecting a leader node for each cluster in the first plurality of clusters, wherein the leader node is coupled at least to a second cluster in the first plurality of clusters;

receiving data at an ingress node, wherein the ingress node is a member of one of the clusters;

transmitting the data to the leader node of the cluster associated with the ingress node;

from the leader node, transmitting the data to an egress node in the MPLS network;

in response to transmitting the data to the egress node, taking a second measurement of the one or more QoS metrics for the plurality of nodes;

regrouping the plurality of nodes into a second plurality of clusters, such that the second plurality of clusters has a hierarchical rank based on the second measurement of the one or more QoS metrics.

2. The method of claim 1, wherein the one or more QoS metrics include one or more of the group consisting of jitter, delay, loss, bandwidth, packet loss ratio.

3. The method of claim 2, wherein the one or more QoS metrics are arranged as a vector.

4. The method of claim of 2, wherein the one or more QoS metrics are ordered lexically.

5 The method of claim 2, wherein the hierarchical ranks based on the first and second measurements of the one or more QoS metrics are based on a weighted average on the one or more QoS metrics.

6. The method of claim 2, wherein the selecting the leader node for each cluster in the first plurality of clusters further includes determining a most-linked node for each cluster in the first plurality of clusters.

7. The method of claim 2, wherein the data includes a threshold for the one or more QoS metrics.

8. The method of claim 2, wherein regrouping the plurality of nodes into a second plurality of clusters further includes merging two or more clusters from the first plurality of clusters.

9. The method of claim 8, wherein the merging to the two or more clusters is in response to deallocating traffic within the two or more clusters.

10. The method of claim 2, wherein regrouping the plurality of nodes into a second plurality of clusters further includes splitting one or more clusters from the first plurality of clusters.

11. The method of claim 10, wherein the splitting one or more clusters from the first plurality of clusters is in response to allocating bandwidth within the one or more clusters.

12. The method of claim 2, wherein regrouping the plurality of nodes into a second plurality of clusters further includes selecting a leader node for each cluster in the second plurality of clusters.

13. The method of claim 2, wherein grouping the nodes into the first plurality of clusters further includes populating a routing table for each node in the network.

14. The method of claim 13, wherein the routing table includes a path to a leader node associated with a next higher cluster in the hierarchy of clusters.

15. The method of claim 13, wherein regrouping the nodes into the second plurality of clusters further includes populating a revised routing table for each node in the network.

16. The method of claim 2, wherein each node in the network includes at least one of a router, a switch, and a terminal.

17 The method of claim 2, wherein the MPLS network includes one or more internetworks.

18. A packet-switched communications network comprising:

plurality of communications nodes;

a plurality of communications links connecting the plurality of communications nodes;

plurality of node clusters, each cluster including one or more nodes from the plurality of communications nodes, wherein the plurality of node clusters are arranged in a hierarchical order defined by one or more QoS parameters, such that for each node cluster, the one or more nodes in the node cluster meet one or more thresholds for the one or more QoS parameters;

wherein the plurality of node clusters gain or lose nodes in response to variations in the one or more QoS parameters.

19. The packet-switched network of claim 18, wherein each cluster of the plurality of node clusters includes a leader node, wherein the leader node has a greatest number of links in the node cluster.

20. The packet-switched network of claim 19, wherein for each cluster, the leader node is in communication with at least one other cluster from the plurality of node clusters.

21. The packet-switched network of claim 20, wherein each node includes a routing table.

22. The packet-switched network of claim 21, wherein for each leader node, the routing table includes a path to all other nodes in the node cluster.

23. The packet-switched network of claim 22, wherein for each leader node, the routing table includes routing information for all node clusters from the plurality of node clusters which have lower rank in the hierarchical order.

24. The packet-switched network of claim 18, wherein the one or more QoS parameters are from the group consisting of latency, jitter, loss, available bandwidth.

25. The packet-switched network of claim 24, wherein the one or more parameters are arranged as a vector.

26. The packet-switched network of claim 24, wherein the one or more QoS parameters are arranged in lexical order.

27. The packet-switched network of claim 24, wherein the one or more QoS parameters are translated to a weighted average.

28. The packet-switched network of claim 18, wherein the traffic is routed between the nodes via the links via Multi Protocol Label Switching.

29. The packet-switched network of claim 18, wherein the packet-switched network is at least partially based on Internet Protocol (IP).

30. The packet-switched network of claim 18, wherein the packet-switched network is at least partially based on Transmission Control Protocol (TCP).

31. The packet-switched network of claim 18, wherein the packet-switched network is at least partially based on Asynchronous Transfer Mode (ATM).

32. The packet-switched network of claim 18, wherein the packet-switched network communicates routing information between the plurality of communications nodes via one or more of the following: Border Gateway Protocol (BGP), Open Shortest Path First (OSPF), Routing Information Protocol (RIP).

33. The packet-switched network of claim 18, wherein the plurality of communications links include Gigabit Ethernet links.

34. A method of routing a packet in a network, the network including a plurality of nodes arranged in a hierarchical plurality of node clusters, the method comprising:

at a first node in the network, determining an egress node for the packet, wherein the first node belongs to a first node cluster from the plurality of node clusters;

determining whether the egress node is within the first node cluster;

if the egress node is not in the first node cluster, transferring the packet to a leader node in the first node cluster;

if the egress node resides in a node cluster which is ranked lower in the hierarchical plurality of node clusters, forwarding the packet to the lower-ranked node cluster;

if the egress node resides outside the first node cluster and the egress node does not reside lower in the hierarchical plurality of node clusters, determining a QoS threshold for the packet;

if the QoS threshold is met by a node cluster ranked higher in the hierarchical plurality of node clusters, forwarding the packet to a lead node of a next higher node cluster in the plurality of node clusters.

35. The method of claim 34, wherein the QoS threshold is contained within a header for the packet.

36. The method of claim 34, wherein the plurality of nodes includes a plurality of routers, switches, and terminals.

37. The method of claim 34, wherein the plurality of nodes communicate at least partially via Internet Protocol (IP).

38. The method of claim 37, wherein the plurality of nodes communicate at least partially via IP version 4.

39. The method of claim 37, wherein the plurality of nodes communicate at least partially via IP version 6.

40. The method of claim 34, wherein the plurality of nodes communicate at least partially via Transmission Control Protocol (TCP).

41. The method of claim 34, wherein the plurality of nodes communicate routing information via one or more of the group consisting of BGP, RIP, and OSPF.

42. The method of claim 34, wherein the network supports Multi Protocol Label Switching.

43. The method of claim 34, wherein the QoS threshold measures one or more of the parameters consisting of delay, jitter, loss, available bandwidth.

44. The method of claim 43, wherein the QoS threshold is a vector representation of the one or more of the parameters.

45. The method of claim 44, wherein the hierarchical plurality of node clusters is arranged according to the one or more parameters.

46. The method of claim 45, wherein the one or more parameters are in lexical order.

47. The method of claim 34, wherein the leader node includes a routing table, the routing table including a path for each node in the first cluster.

48. The method of claim 47, wherein the routing table includes a path for each node in the node cluster which is ranked lower in the hierarchical plurality of node clusters.

49. The method of claim 48, wherein forwarding the packet to the lower-ranked node cluster further includes identifying the egress node in the routing table for the leader node in the first cluster.

50. A method of populating a routing table for a communications node in a network, wherein the network is organized as a plurality of hierarchically-arranged node clusters, such that the node is a member of a node cluster in the plurality of node clusters, the method comprising:

determining a shortest path to each node in a plurality of other nodes within the cluster;

entering the shortest path to each of the plurality of other nodes within the cluster in the routing table;

determining a minimum bandwidth link for each node in the plurality of other nodes in the cluster;

entering the minimum bandwidth link in the routing table;

selecting exactly one leader node for the cluster, selecting the leader node further including selecting a node from the node cluster with a greatest number of links to the plurality of other nodes;

in the leader node, determining a shortest path to each node in a plurality of node clusters of lower rank, and entering the shortest path to each node in the lower ranked clusters in the routing table.

51. The method of claim 50, wherein selecting exactly one leader node further includes:

if there is more than one node in the node cluster with the greatest number of links, summing the minimum bandwidth link for each of the more than one nodes.

52. The method of claim 50, further including:

in the leader node, calculating a shortest path to a leader node of a node cluster of next highest rank.

53. The method of claim 50, wherein determining the shortest path to each node in a plurality of other nodes within the cluster further includes applying Dijkstra's Shortest Path algorithm.

54. The method of claim 50, wherein the network is at least partially an MPLS-based network.

55. The method of claim 50, wherein the network is at least partially an IP-based network.

56. The method of claim 50, wherein the network is at least partially a TCP-based network.

57. The method of claim 50, wherein the plurality of hierarchically arranged node clusters are ranked according to one or more QoS parameters for the network.

58. The method of claim 50, wherein the QoS parameters include one or more of the group consisting of jitter, delay, loss, available bandwidth.

59. The method of claim 50, wherein the network is at least partially an ATM network.

60. The method of claim 50, wherein the network communicates routing information by use of one or more of the group consisting of RIP, OSPF, BGP.

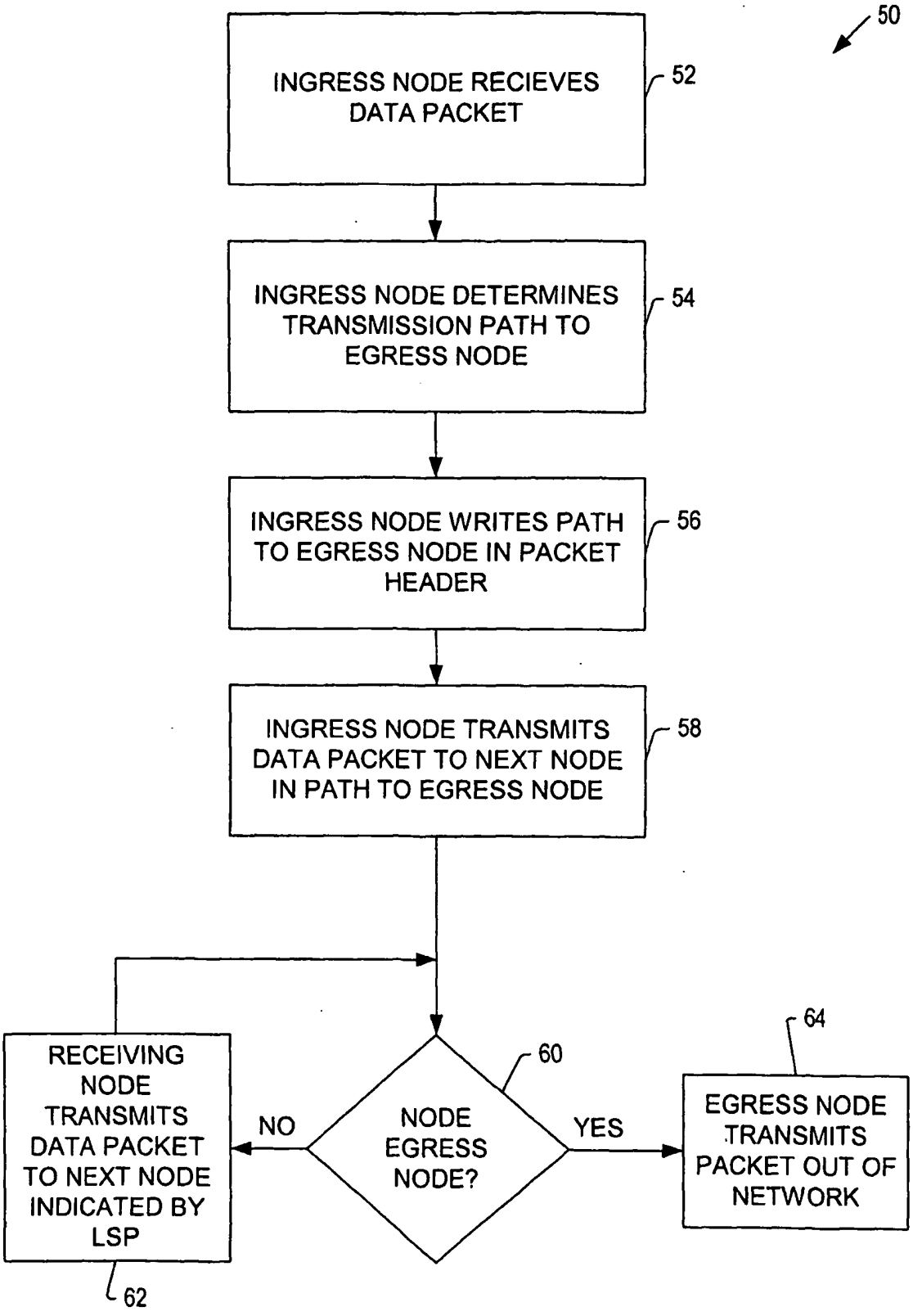


FIG. 1
PRIOR ART

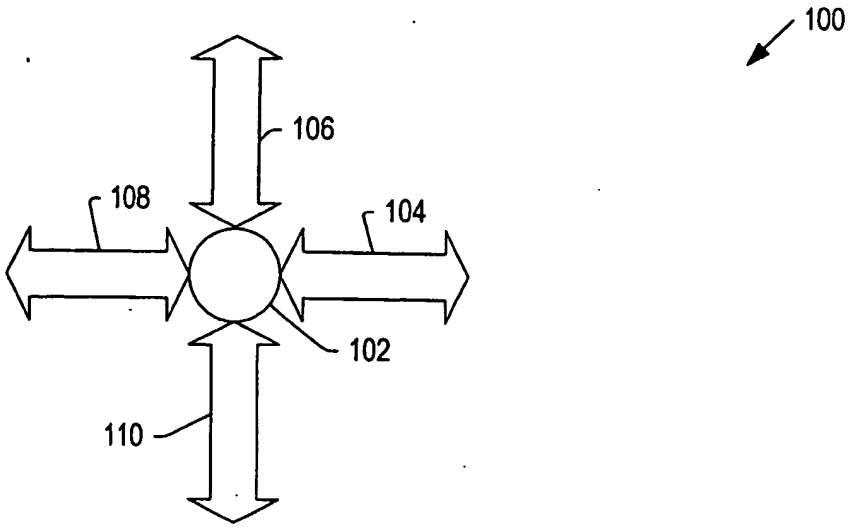
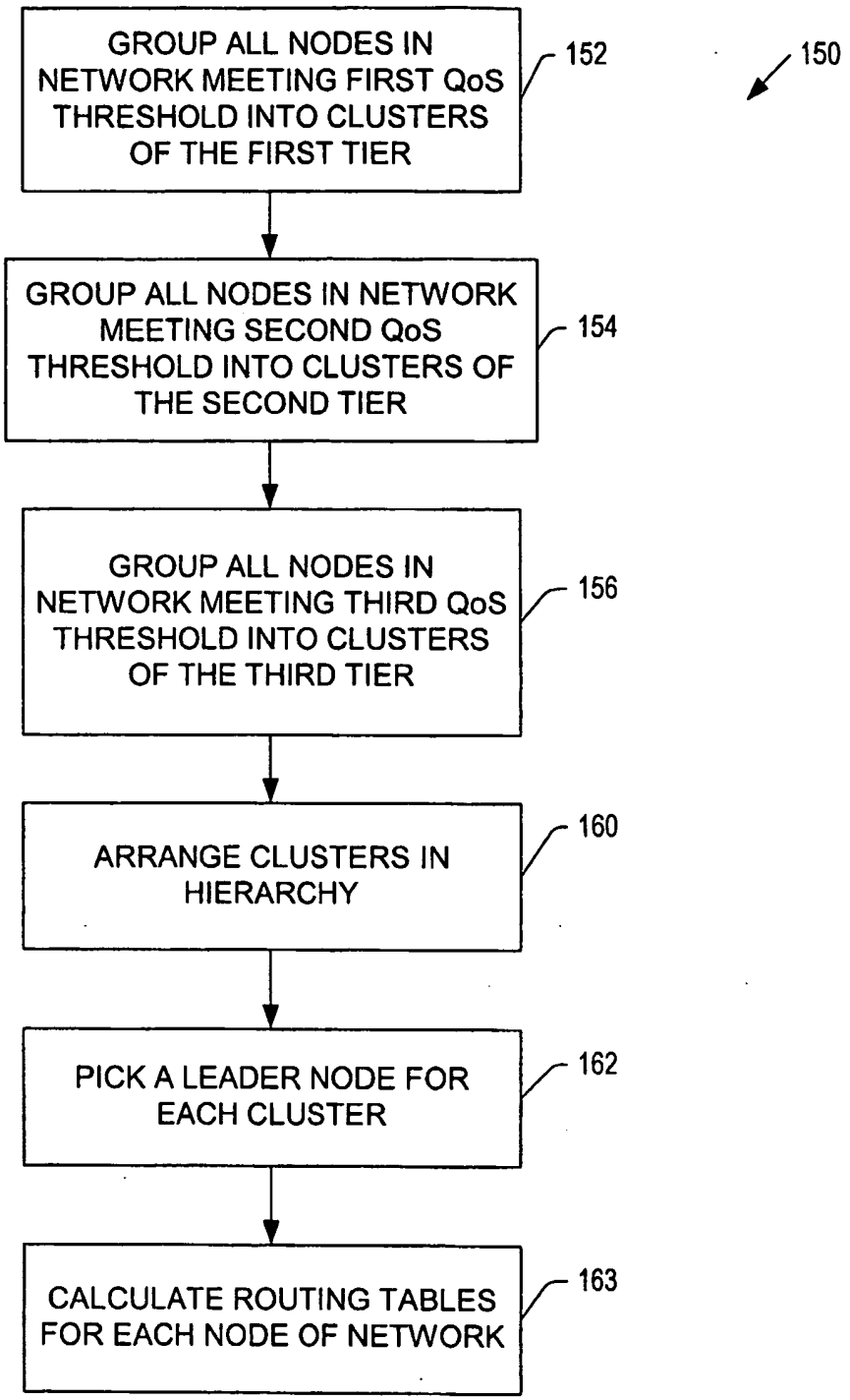


FIG. 2


GROUP ALL NODES IN NETWORK MEETING FIRST QoS THRESHOLD INTO CLUSTERS OF THE FIRST TIER
152
150
GROUP ALL NODES IN NETWORK MEETING SECOND QoS THRESHOLD INTO CLUSTERS OF THE SECOND TIER
154
GROUP ALL NODES IN NETWORK MEETING THIRD QoS THRESHOLD INTO CLUSTERS OF THE THIRD TIER
156
ARRANGE CLUSTERS IN HIERARCHY
160
PICK A LEADER NODE FOR EACH CLUSTER
162
CALCULATE ROUTING TABLES FOR EACH NODE OF NETWORK
163

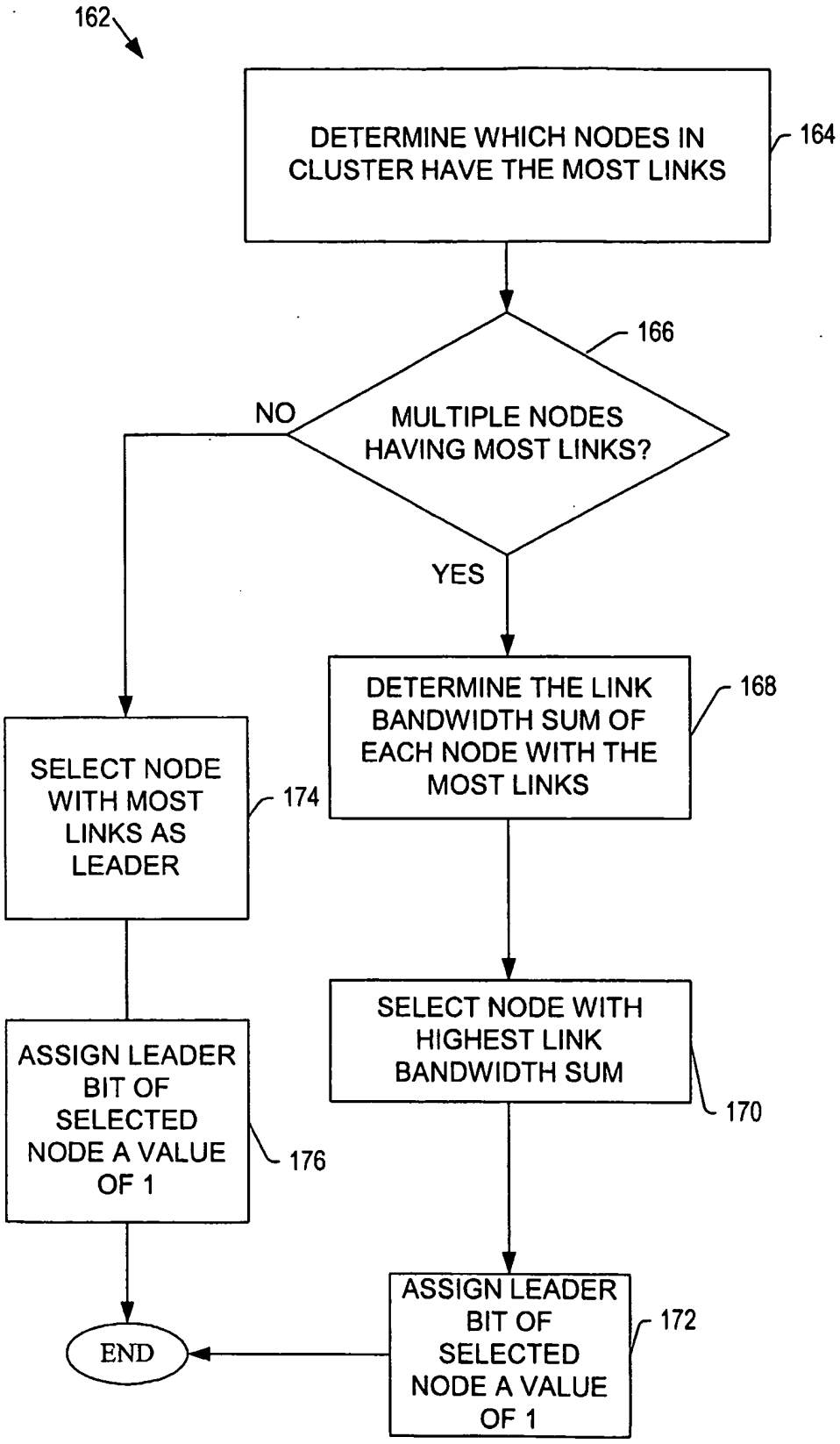

162
DETERMINE WHICH NODES IN CLUSTER HAVE THE MOST LINKS
164
166
MULTIPLE NODES HAVING MOST LINKS?
NO
YES
DETERMINE THE LINK BANDWIDTH SUM OF EACH NODE WITH THE MOST LINKS
168
SELECT NODE WITH MOST LINKS AS LEADER
174
SELECT NODE WITH HIGHEST LINK BANDWIDTH SUM
170
ASSIGN LEADER BIT OF SELECTED NODE A VALUE OF 1
176
ASSIGN LEADER BIT OF SELECTED NODE A VALUE OF 1
172
END

FIG. 4

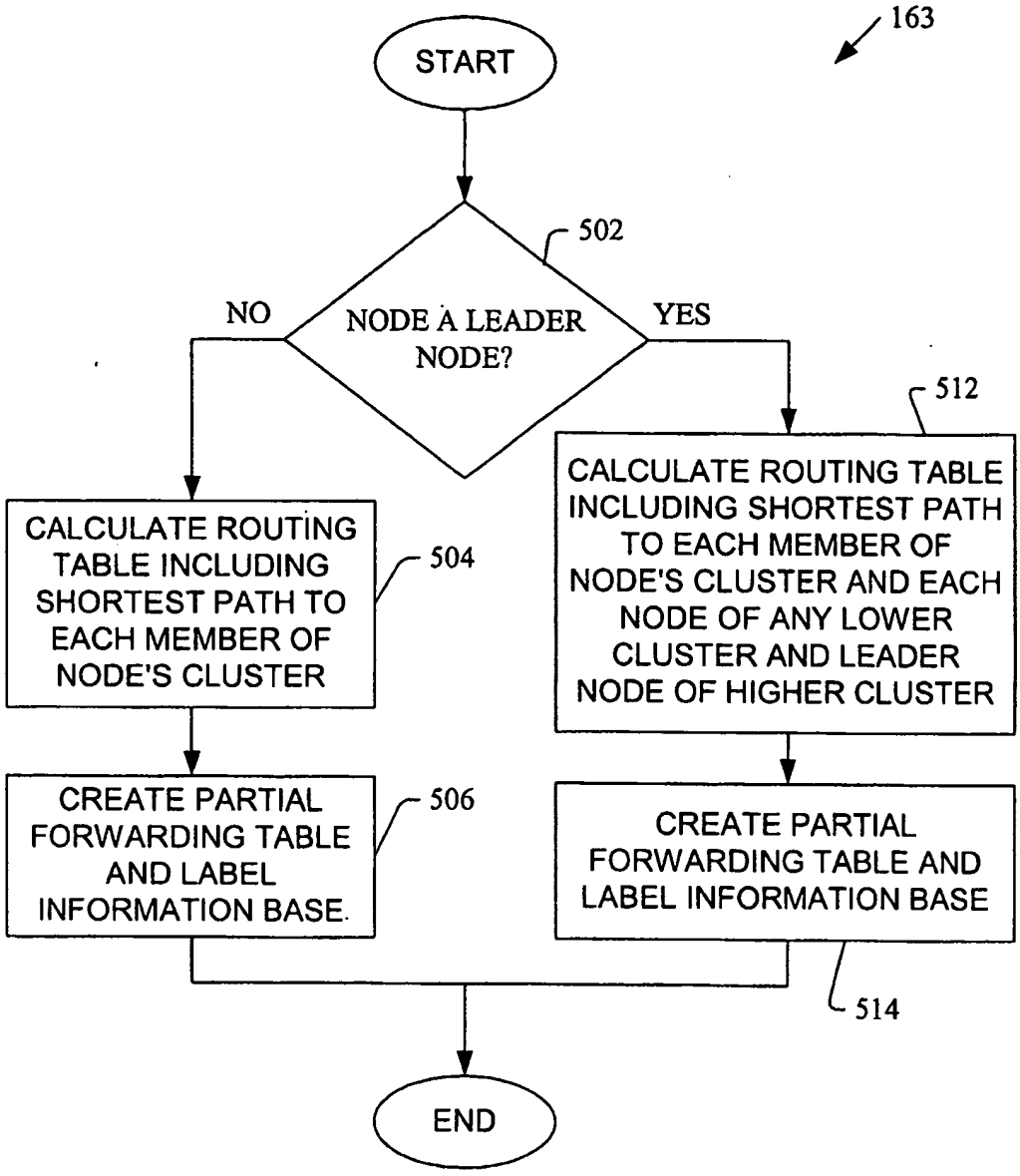

163
START
502
NODE A LEADER NODE?
NO
YES
CALCULATE ROUTING TABLE INCLUDING SHORTEST PATH TO EACH MEMBER OF NODE'S CLUSTER
504
CREATE PARTIAL FORWARDING TABLE AND LABEL INFORMATION BASE.
506
512
CALCULATE ROUTING TABLE INCLUDING SHORTEST PATH TO EACH MEMBER OF NODE'S CLUSTER AND EACH NODE OF ANY LOWER CLUSTER AND LEADER NODE OF HIGHER CLUSTER
CREATE PARTIAL FORWARDING TABLE AND LABEL INFORMATION BASE
514
END

FIG. 5

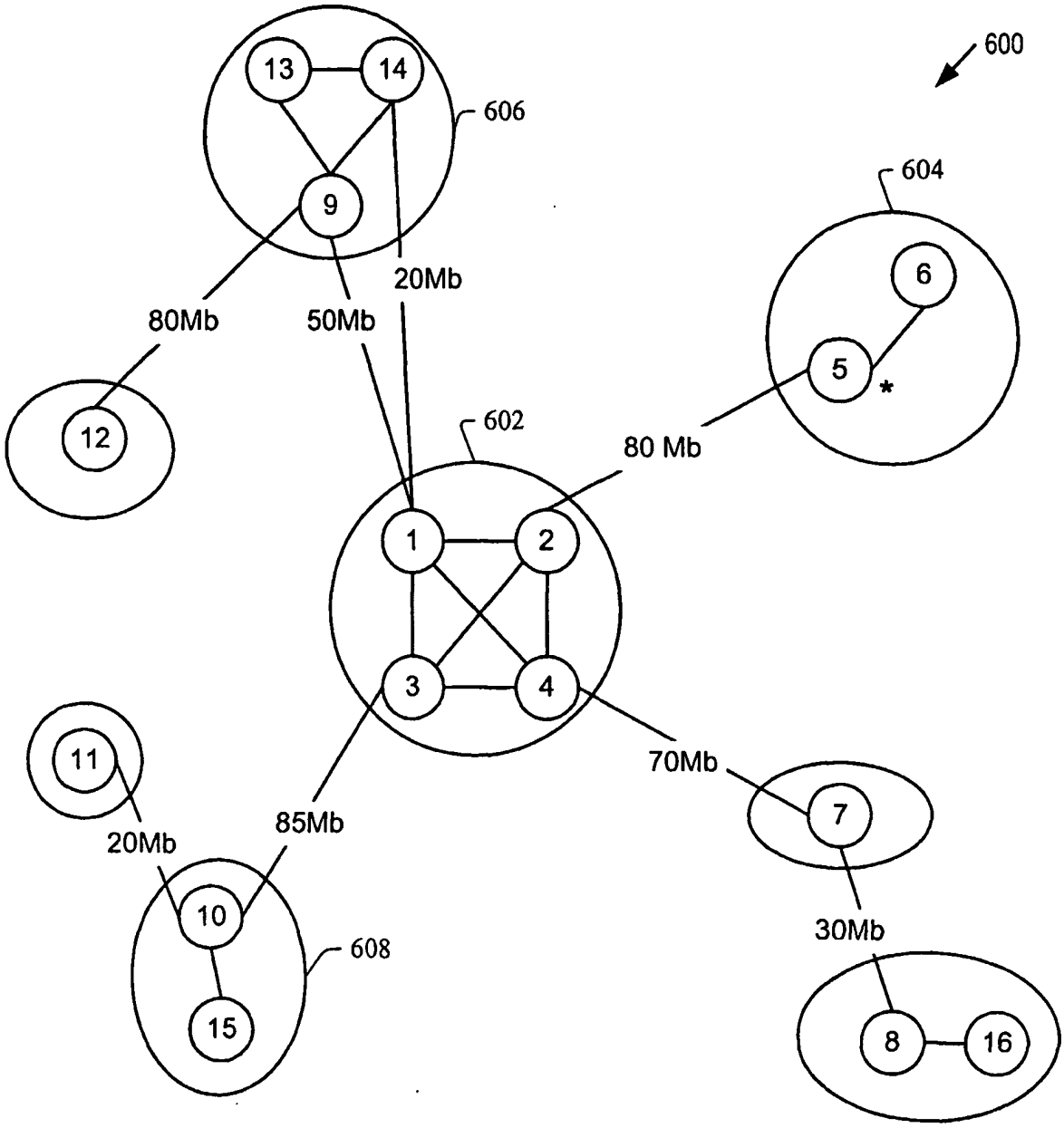

600
606
13
14
9
604
6
5
*
20Mb
80Mb
50Mb
12
602
80 Mb
1
2
3
4
11
70Mb
7
85Mb
20Mb
10
608
30Mb
15
8
16

FIG. 6A

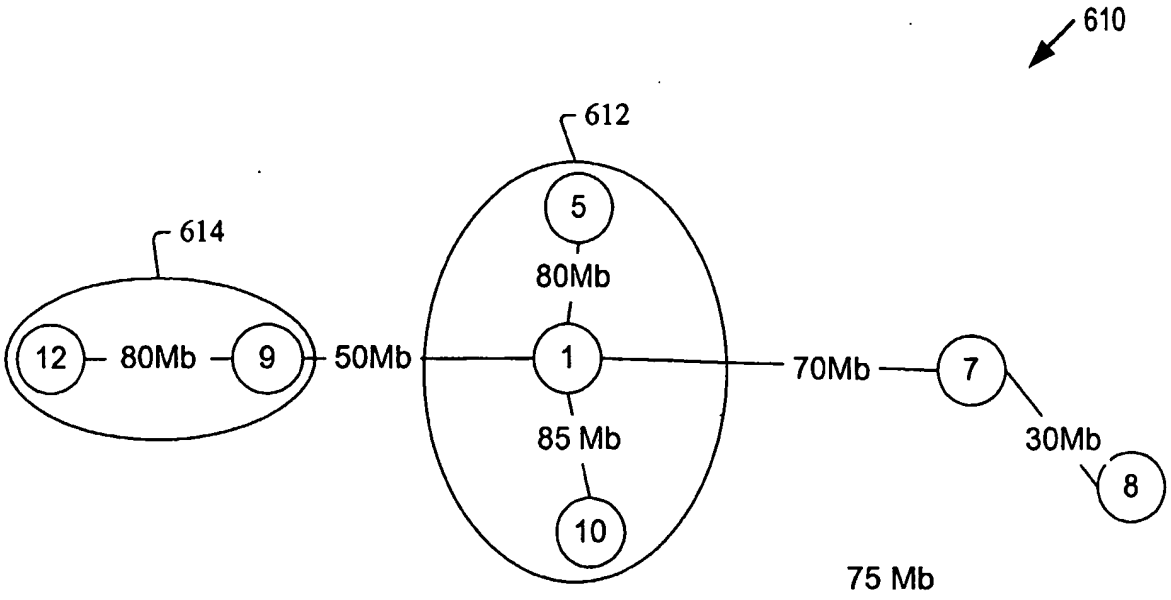

610
612
5
80Mb
614
12
80Mb
9
50Mb
1
70Mb
7
30Mb
8
85 Mb
10
75 Mb

FIG. 6B

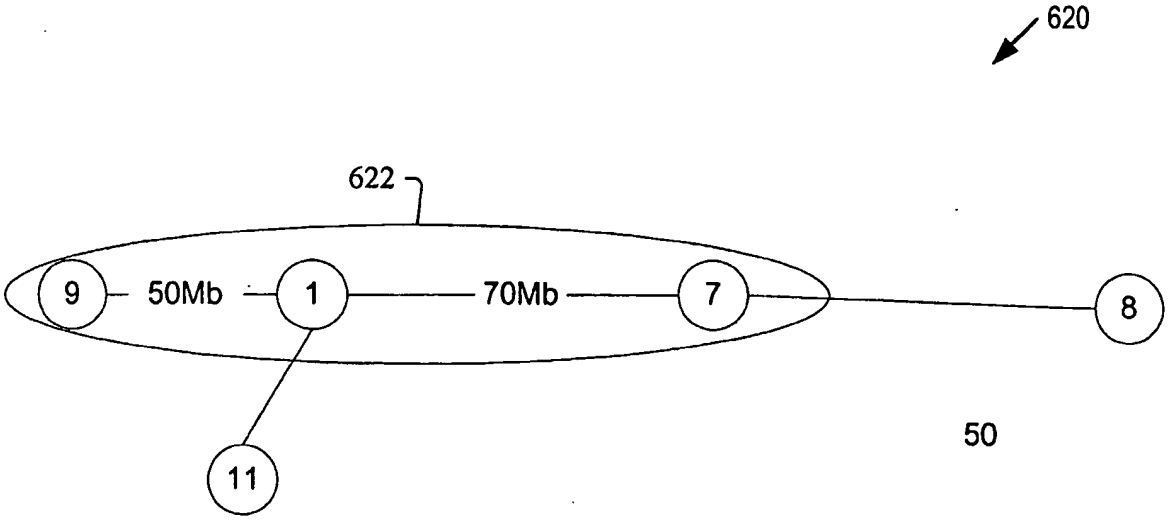

620
622
9
50Mb
1
70Mb
7
8
11
50

FIG. 6C

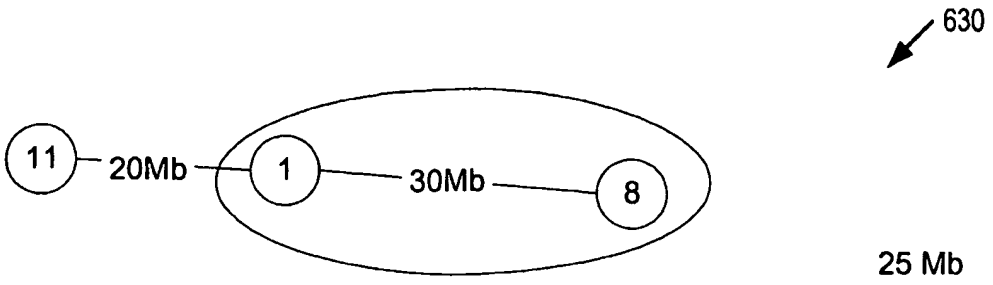


FIG. 6D

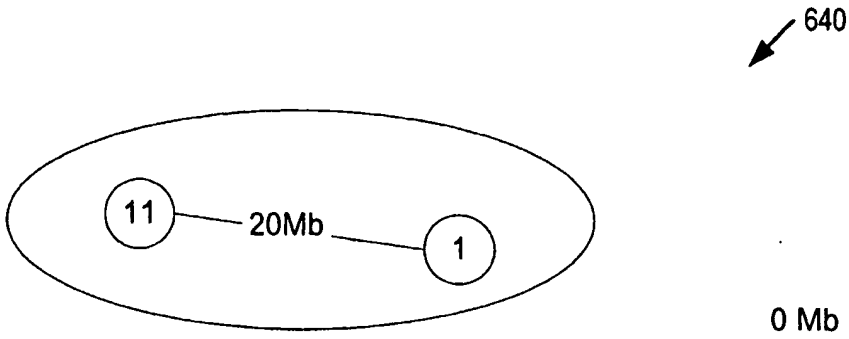


FIG. 6E

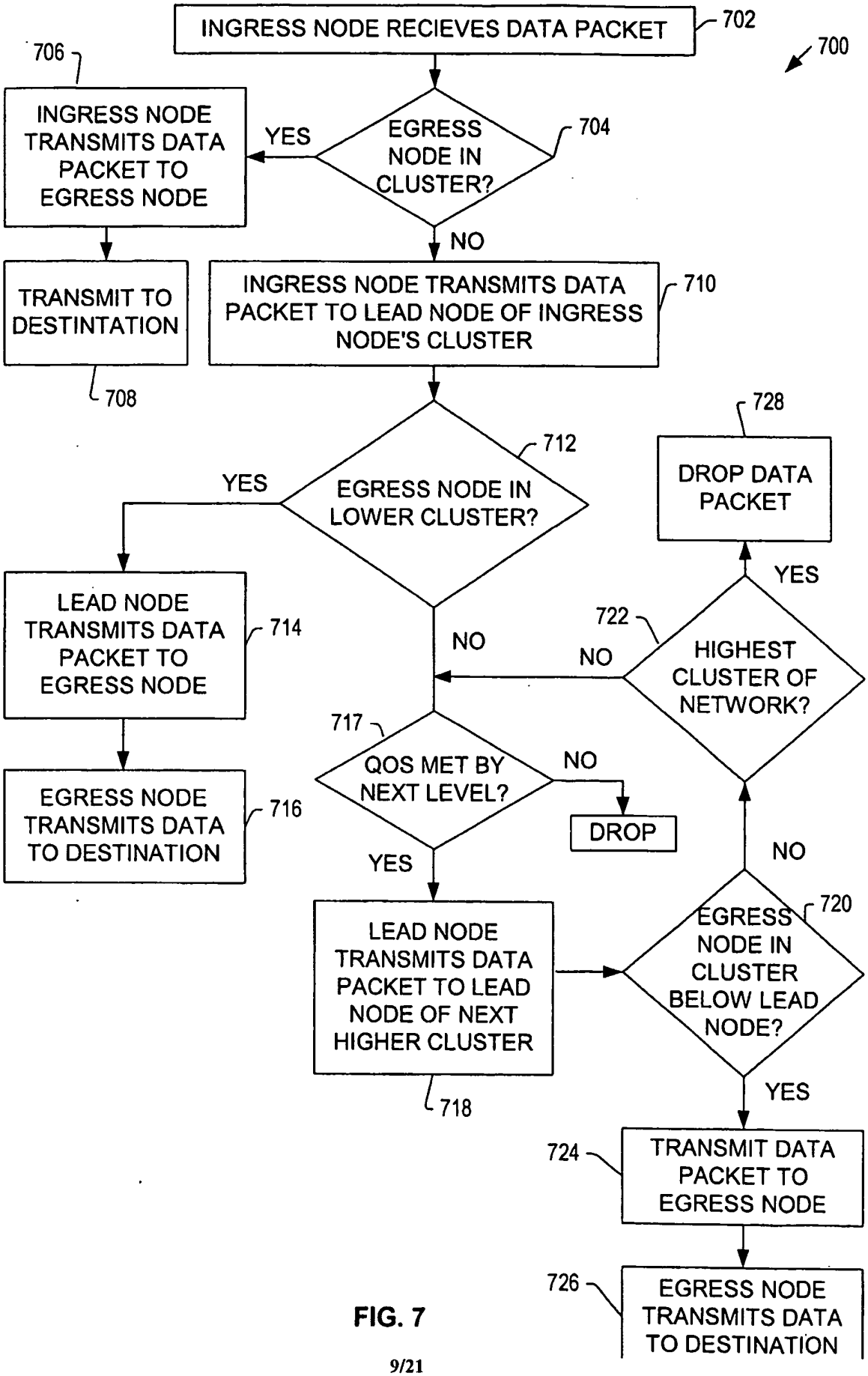

INGRESS NODE RECIEVES DATA PACKET
702
700
706
INGRESS NODE TRANSMITS DATA PACKET TO EGRESS NODE
YES
EGRESS NODE IN CLUSTER?
704
NO
TRANSMIT TO DESTINTATION
708
INGRESS NODE TRANSMITS DATA PACKET TO LEAD NODE OF INGRESS NODE'S CLUSTER
710
728
DROP DATA PACKET
712
YES
EGRESS NODE IN LOWER CLUSTER?
YES
LEAD NODE TRANSMITS DATA PACKET TO EGRESS NODE
714
722
NO
NO
HIGHEST CLUSTER OF NETWORK?
717
QOS MET BY NEXT LEVEL?
NO
DROP
EGRESS NODE TRANSMITS DATA TO DESTINATION
716
YES
NO
LEAD NODE TRANSMITS DATA PACKET TO LEAD NODE OF NEXT HIGHER CLUSTER
720
EGRESS NODE IN CLUSTER BELOW LEAD NODE?
718
YES
724
TRANSMIT DATA PACKET TO EGRESS NODE
726
EGRESS NODE TRANSMITS DATA TO DESTINATION

FIG. 7

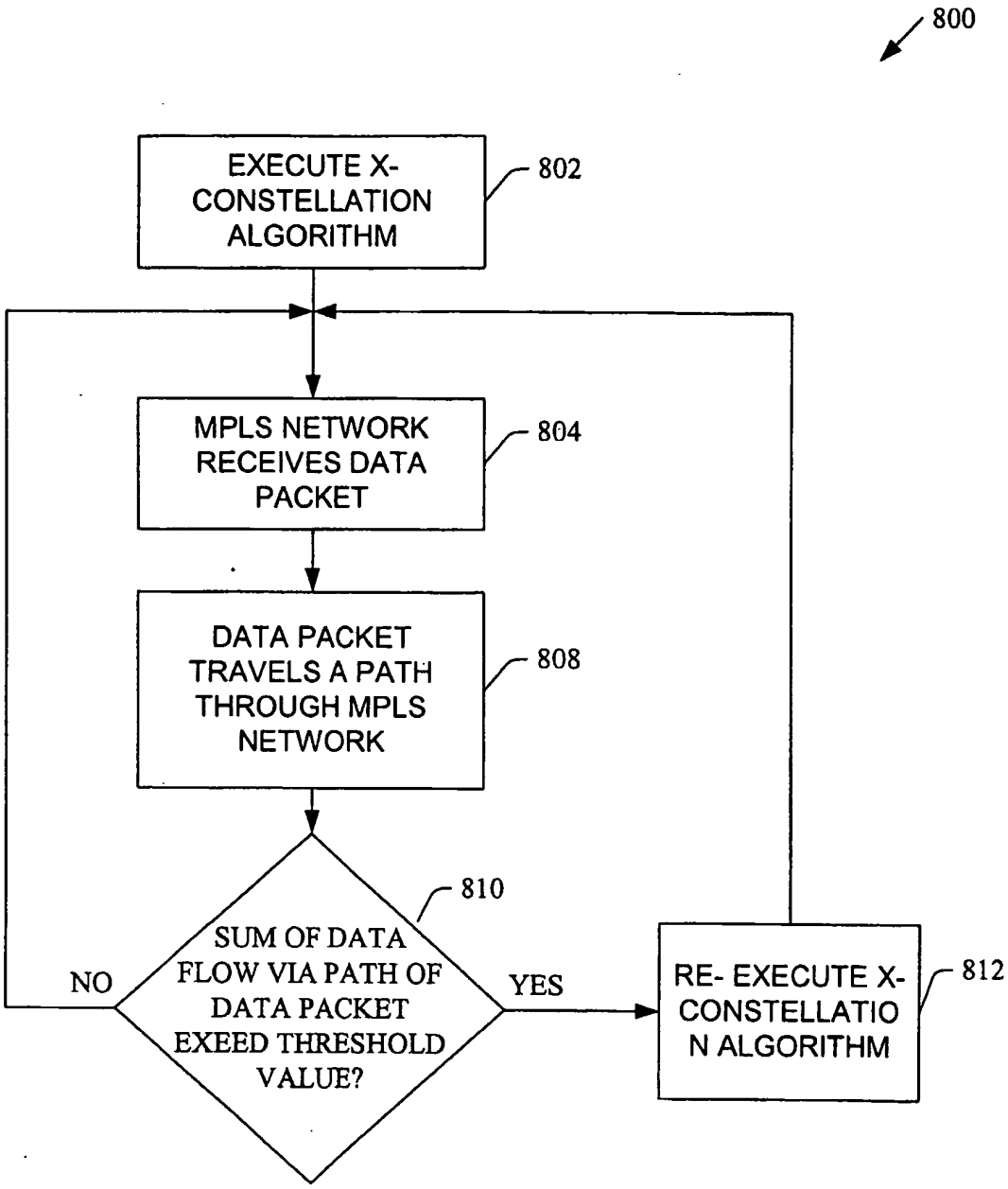

800
EXECUTE X-CONSTELLATION ALGORITHM
802
MPLS NETWORK RECEIVES DATA PACKET
804
DATA PACKET TRAVELS A PATH THROUGH MPLS NETWORK
808
SUM OF DATA FLOW VIA PATH OF DATA PACKET EXEED THRESHOLD VALUE?
810
NO
YES
RE- EXECUTE X-CONSTELLATIO N ALGORITHM
812

FIG. 8

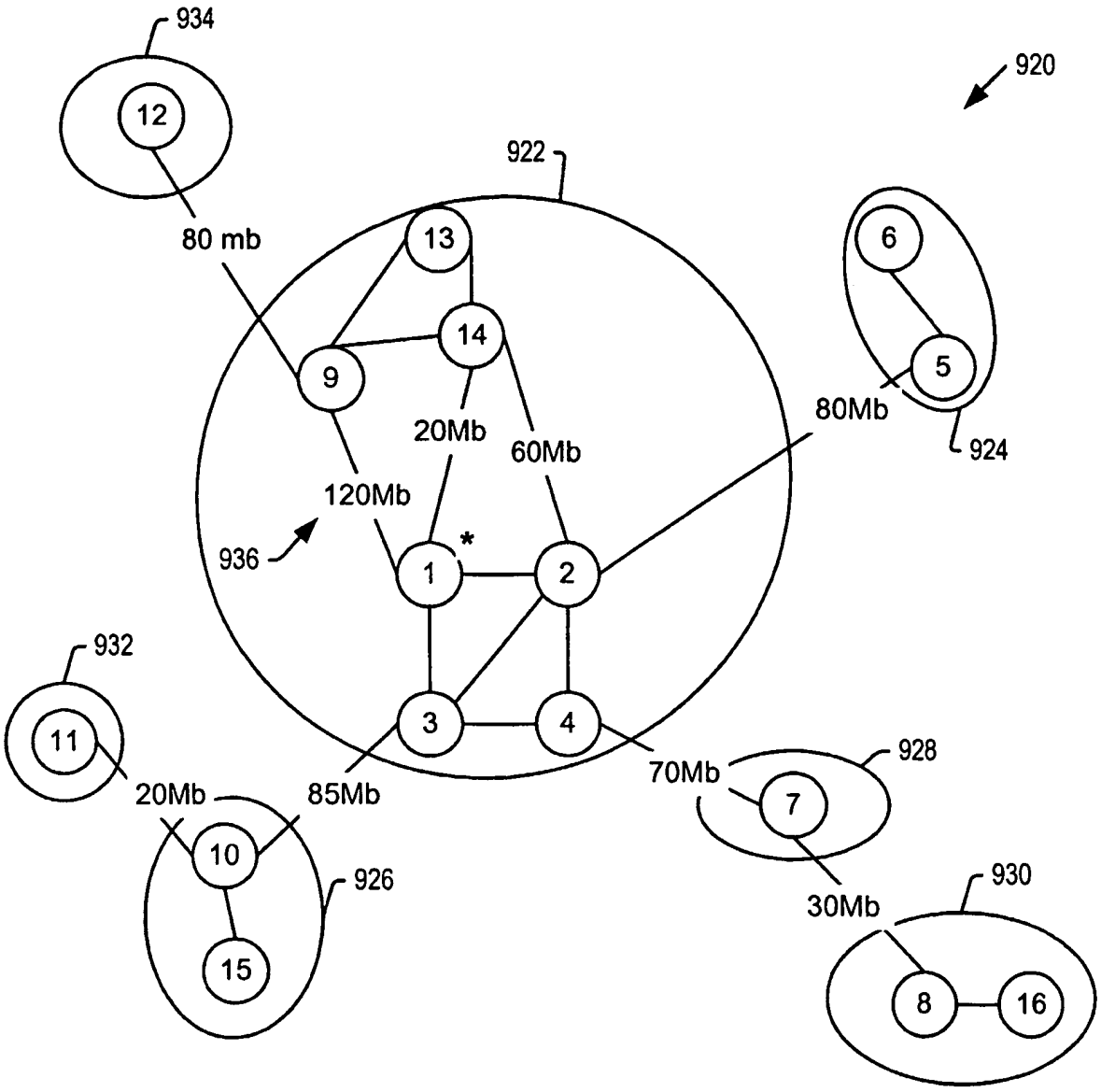

934
12
920
922
80 mb
13
6
14
9
5
80Mb
20Mb
60Mb
924
120Mb
936
*
1
2
932
11
3
4
70Mb
928
20Mb
85Mb
7
10
926
930
30Mb
15
8
16

FIG. 9A

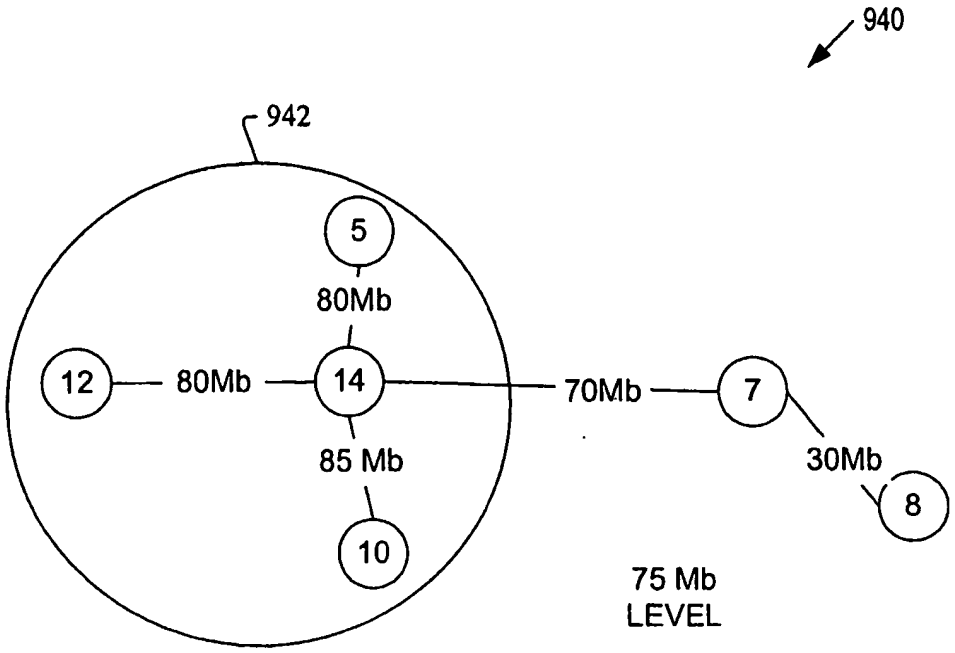

940
942
5
80Mb
12
80Mb
14
70Mb
7
30Mb
8
85 Mb
10
75 Mb
LEVEL

FIG. 9B

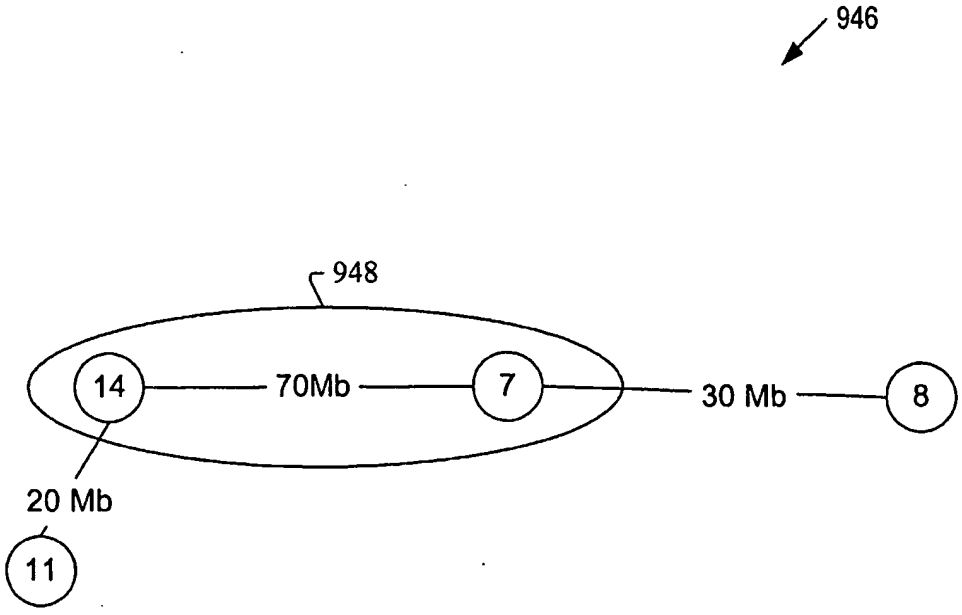

946
948
14
70Mb
7
30 Mb
8
20 Mb
11

FIG. 9C

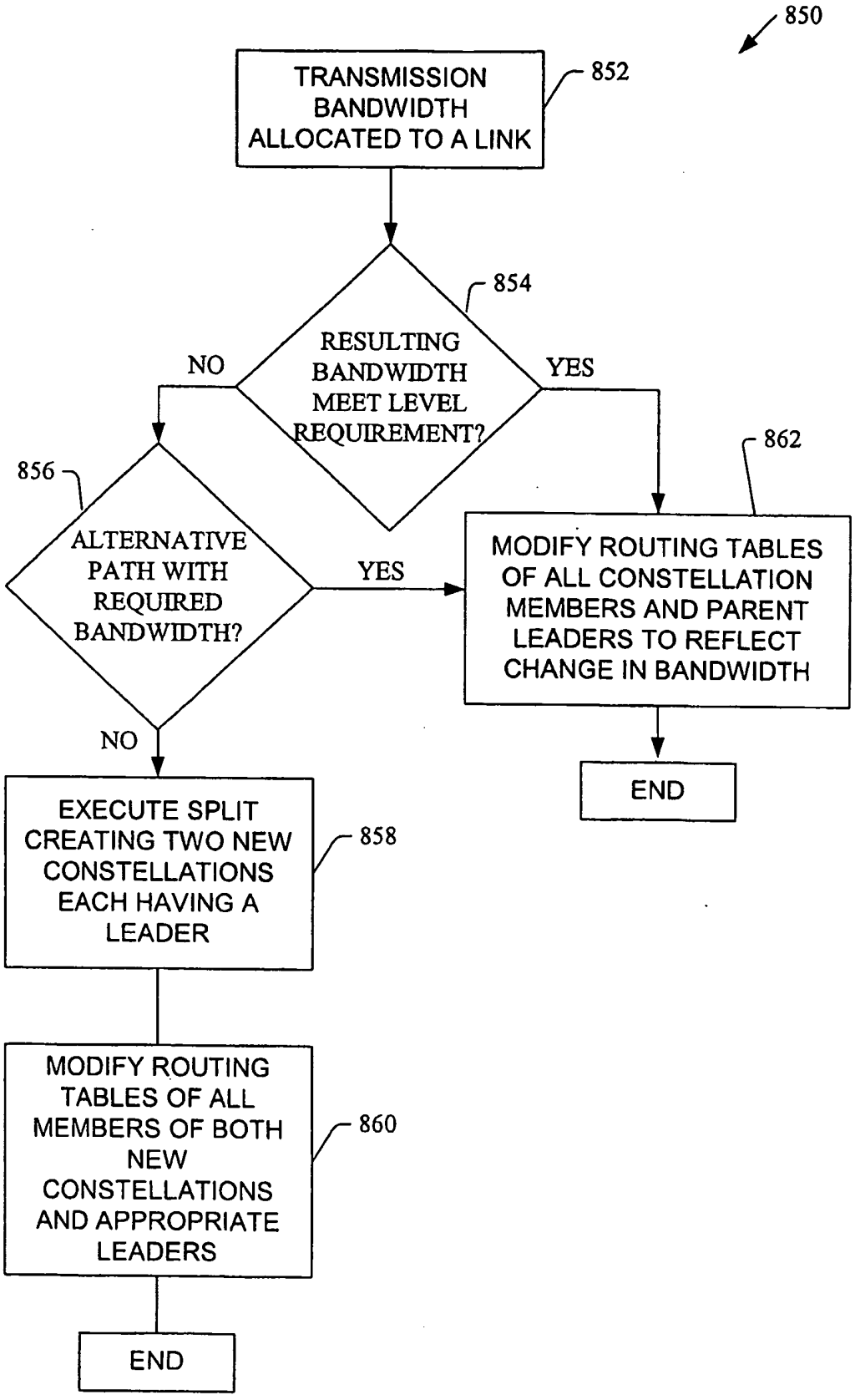

850
TRANSMISSION BANDWIDTH ALLOCATED TO A LINK
852
RESULTING BANDWIDTH MEET LEVEL REQUIREMENT?
854
NO
YES
ALTERNATIVE PATH WITH REQUIRED BANDWIDTH?
856
YES
NO
MODIFY ROUTING TABLES OF ALL CONSTELLATION MEMBERS AND PARENT LEADERS TO REFLECT CHANGE IN BANDWIDTH
862
END
EXECUTE SPLIT CREATING TWO NEW CONSTELLATIONS EACH HAVING A LEADER
858
MODIFY ROUTING TABLES OF ALL MEMBERS OF BOTH NEW CONSTELLATIONS AND APPROPRIATE LEADERS
860
END

FIG. 9D

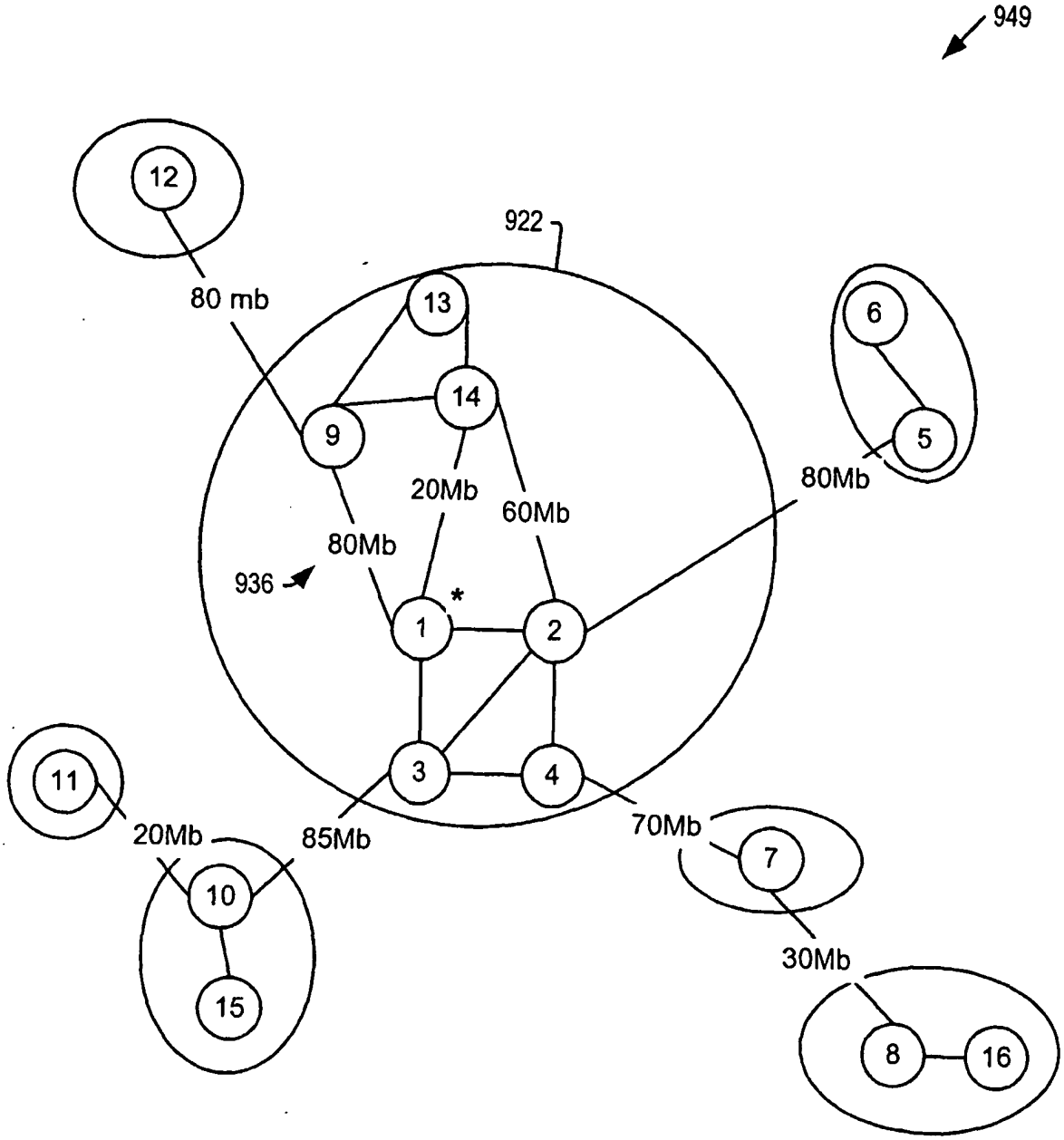

949
922
12
80 mb
13
14
9
20Mb
60Mb
80Mb
936
*
1
2
6
5
80Mb
11
20Mb
85Mb
3
4
70Mb
7
30Mb
10
15
8
16

FIG. 9E

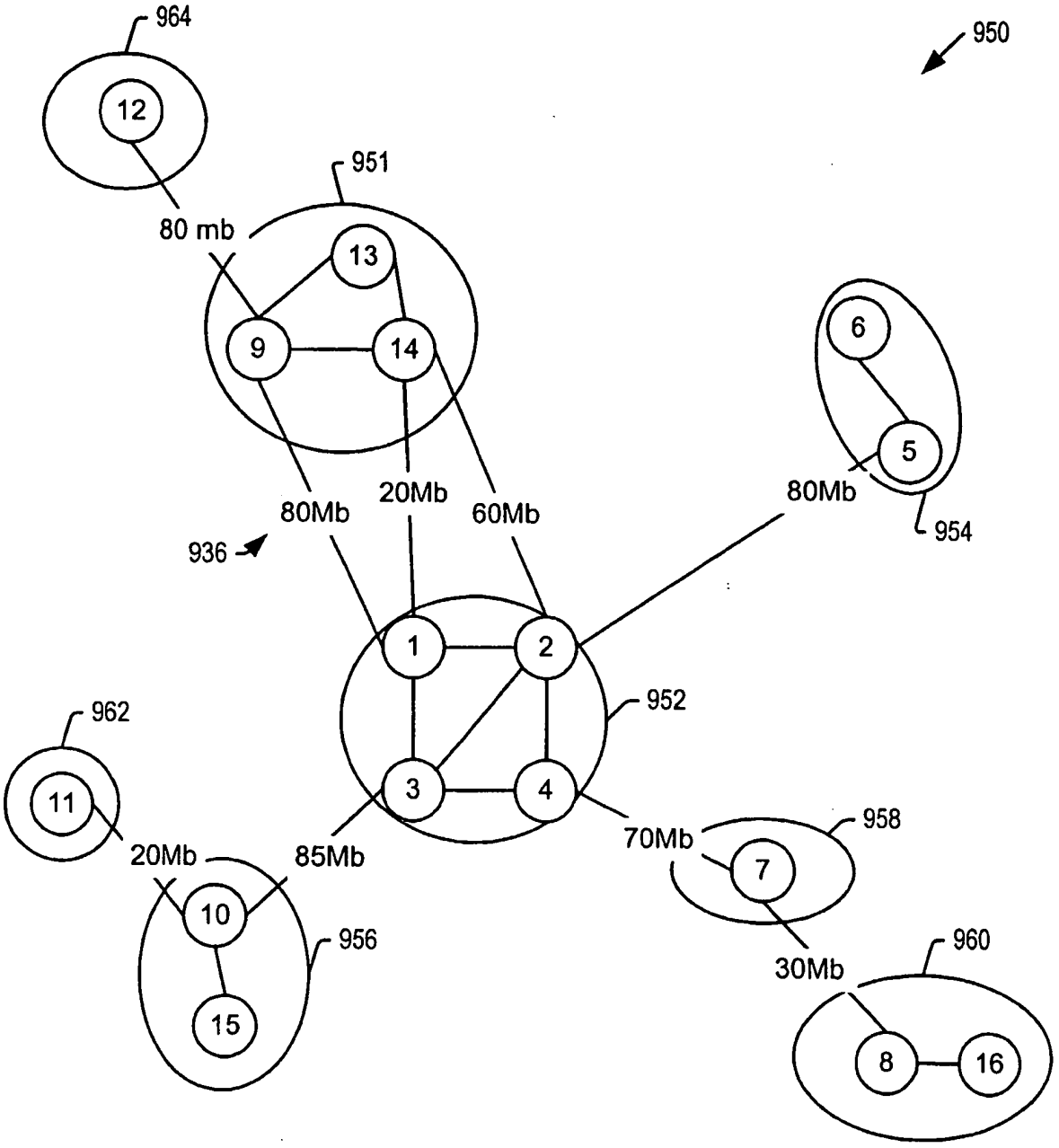

964
950
12
951
80 mb
13
9
14
6
5
80Mb
954
80Mb
20Mb
60Mb
936
1
2
952
962
11
3
4
958
70Mb
7
20Mb
85Mb
10
956
15
30Mb
960
8
16

FIG. 9F

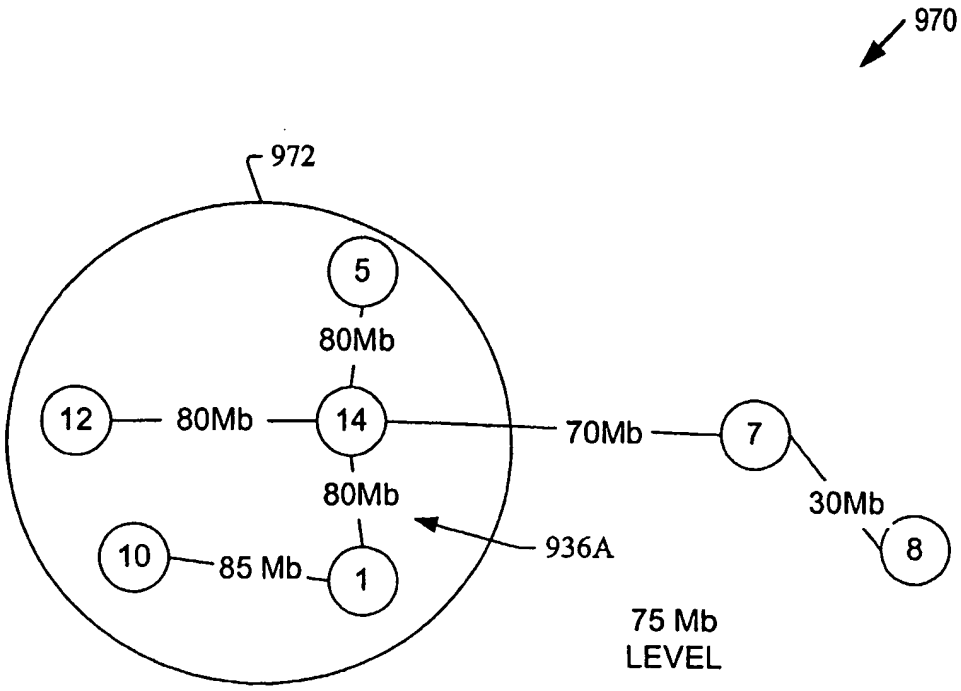

970
972
5
80Mb
12
80Mb
14
70Mb
7
80Mb
30Mb
936A
10
85 Mb
1
8
75 Mb
LEVEL

FIG. 9G

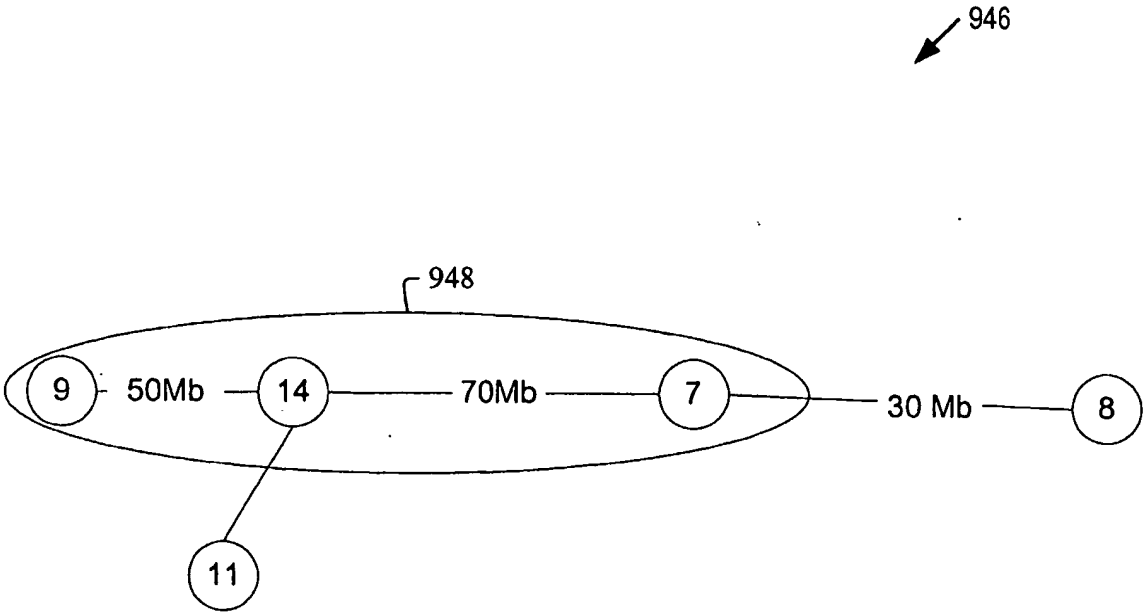

946
948
9
50Mb
14
70Mb
7
30 Mb
8
11

FIG. 9H

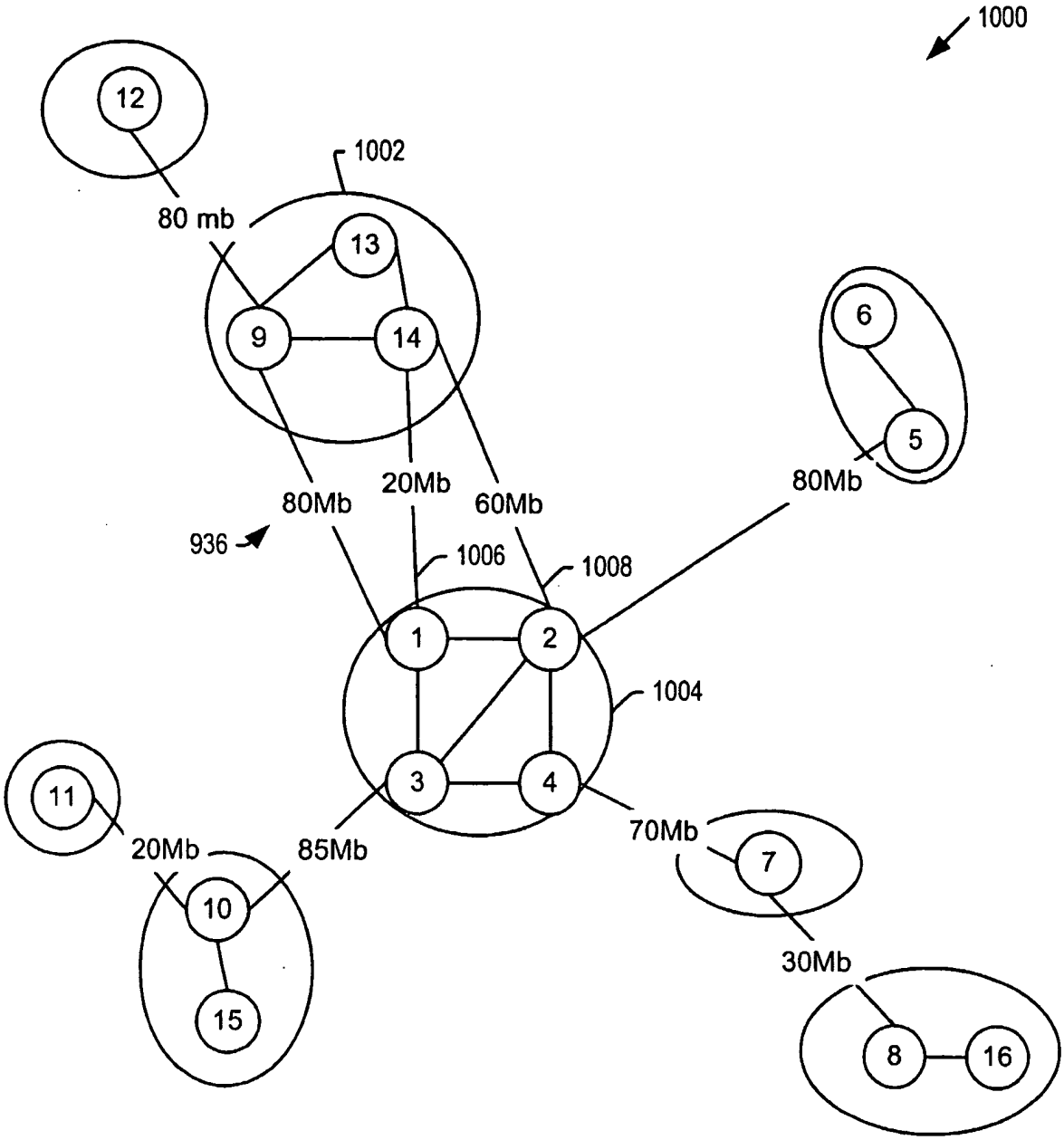

1000
12
1002
80 mb
13
9
14
6
5
80Mb
20Mb
60Mb
80Mb
936
1006
1008
1
2
1004
3
4
11
20Mb
85Mb
70Mb
7
10
30Mb
15
8
16

FIG. 10A

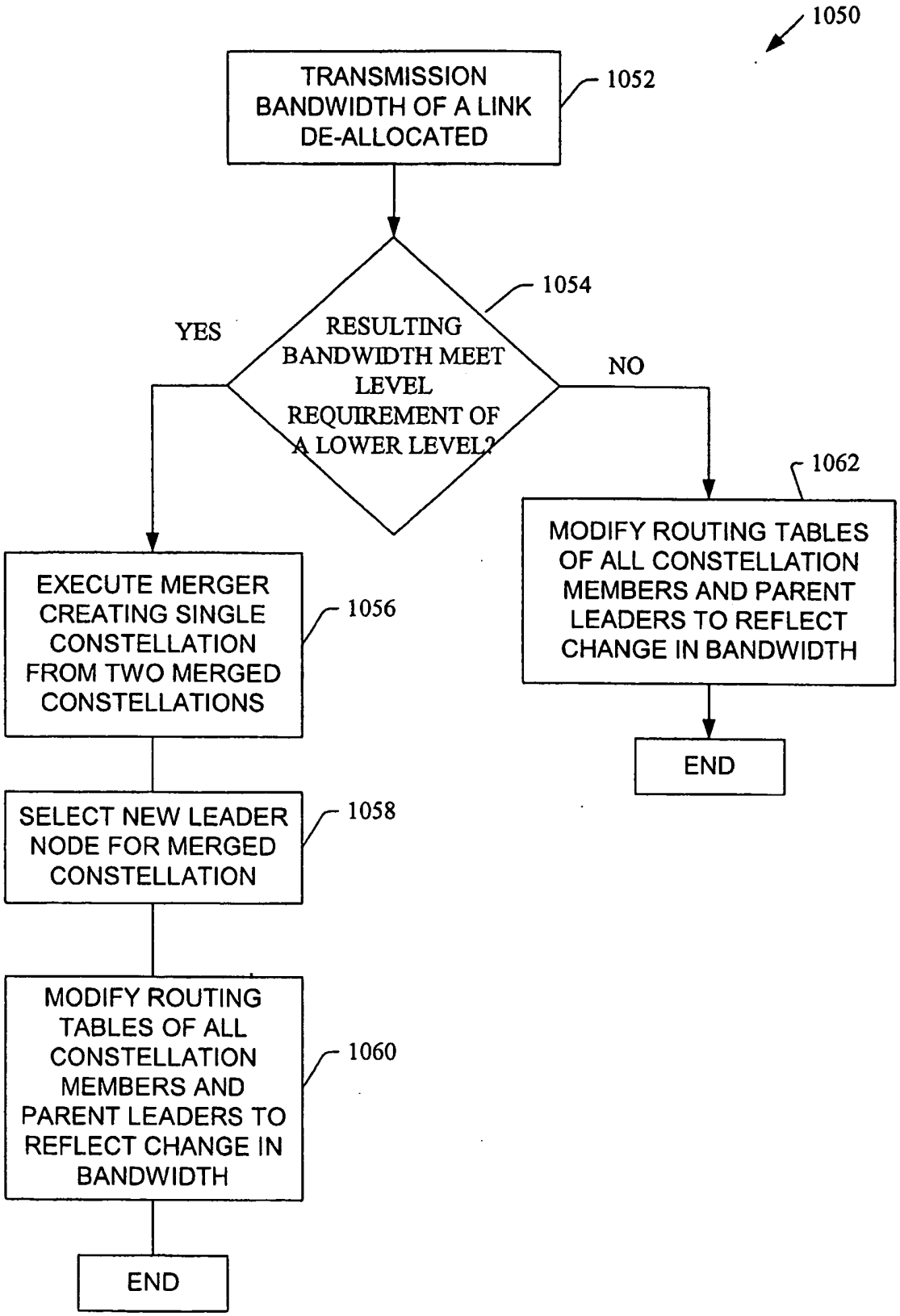

1050
TRANSMISSION BANDWIDTH OF A LINK DE-ALLOCATED
1052
RESULTING BANDWIDTH MEET LEVEL REQUIREMENT OF A LOWER LEVEL?
1054
YES
NO
EXECUTE MERGER CREATING SINGLE CONSTELLATION FROM TWO MERGED CONSTELLATIONS
1056
SELECT NEW LEADER NODE FOR MERGED CONSTELLATION
1058
MODIFY ROUTING TABLES OF ALL CONSTELLATION MEMBERS AND PARENT LEADERS TO REFLECT CHANGE IN BANDWIDTH
1060
END
1062
MODIFY ROUTING TABLES OF ALL CONSTELLATION MEMBERS AND PARENT LEADERS TO REFLECT CHANGE IN BANDWIDTH
END

FIG. 10B

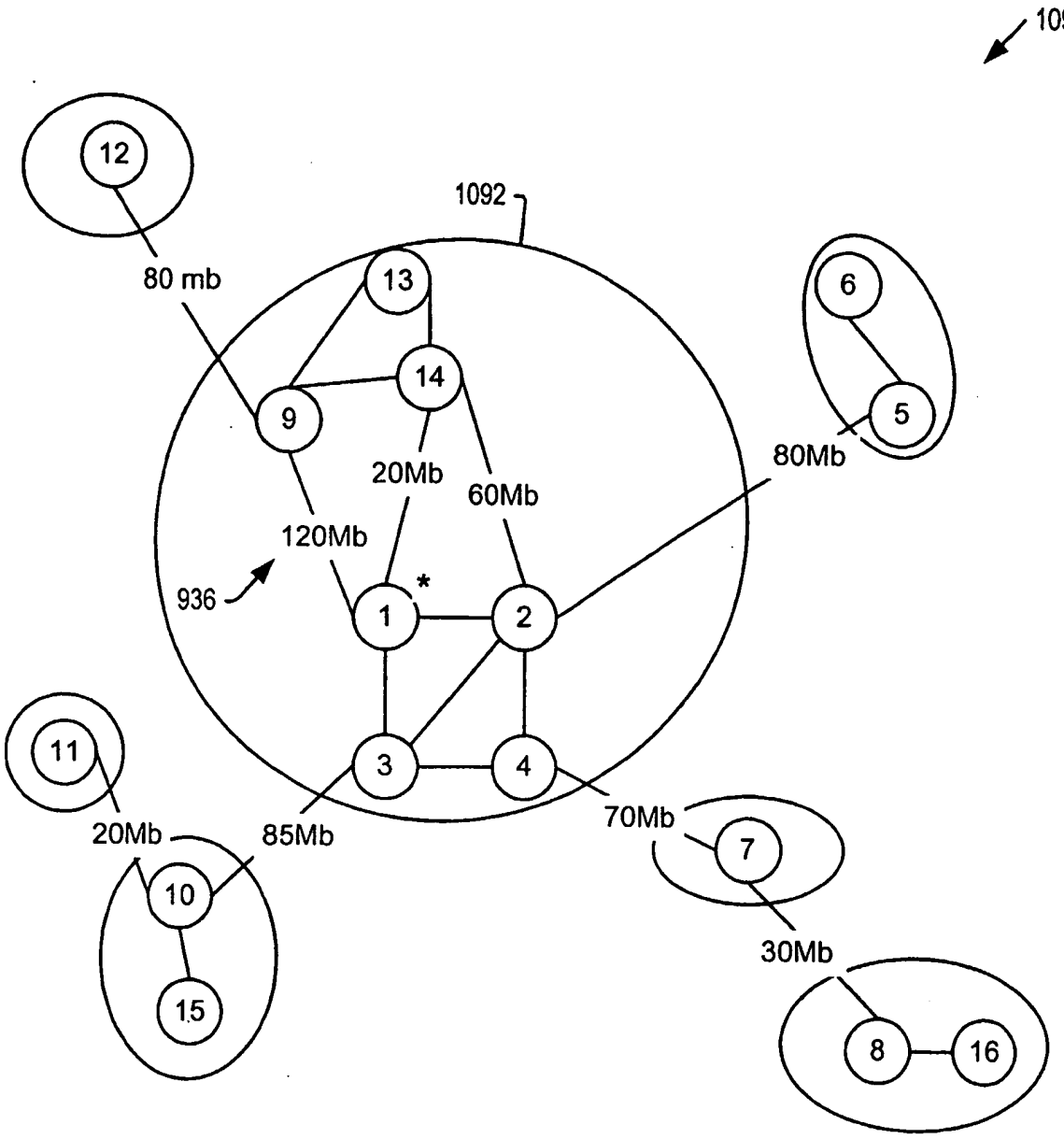

1090
1092
12
80 mb
13
14
9
20Mb
60Mb
120Mb
936
*
1
2
6
5
80Mb
11
20Mb
85Mb
3
4
70Mb
7
30Mb
10
15
8
16

FIG. 10C

1100

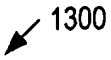
DESTINATION	ROUTE			
NODE 1	FIRST HOP	SECOND	THIRD	HOP X
NODE 2	FIRST HOP	SECOND		
NODE 3	FIRST HOP			
NODE 4	FIRST HOP			
NODE 5	FIRST HOP			
NODE X	FIRST HOP			

1200

Partial Forwarding Table –

FEC	PHB	LIBptr	Alternative Path

FIG. 11



Label Information Base -

iIface	iLabel	oIface	oLabel	LIBptr

FIG. 12

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 July 2003 (17.07.2003)

PCT

(10) International Publication Number
WO 2003/058868 A3

(51) International Patent Classification[7]: H04L 12/66, 12/28

(21) International Application Number: PCT/US2003/000163

(22) International Filing Date: 3 January 2003 (03.01.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/345,834	4 January 2002 (04.01.2002)	US
60/384,438	31 May 2002 (31.05.2002)	US

(71) Applicant *(for all designated States except US)*: **EINFINITUS TECHNOLOGIES, INC.** [IN/IN]; Harshada"A", 127/2, Mahaganesh Colony, Paud Road, Kothrud, Pune 411029, Maharashtra (IN).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **TANDON, Siddharth** [IN/US]; 870 East El Camino real, # 321, Sunnyvale, CA 94087 (US). **BANSAL, Jayant** [IN/IN]; Harshada"A", 127/2, Mahaganesh Colony, Paud Road, Kothrud, Pune 411029, Maharashtra (IN).

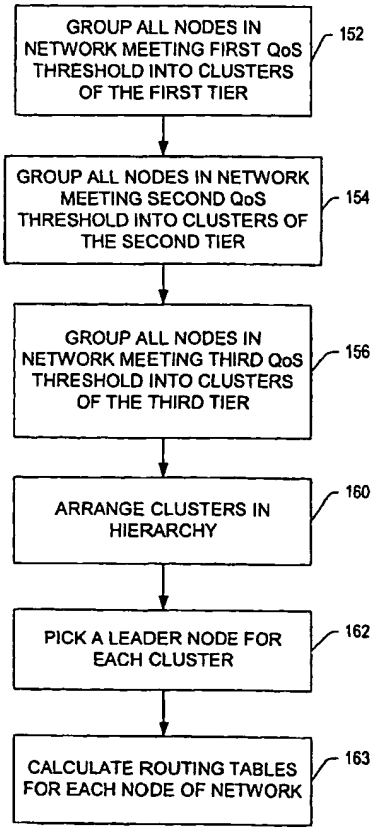
(74) Agent: **COLEMAN, Brian, R.**; Perkins Coie LLP, 101 Jefferson Drive, Menlo Park, CA 94025 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW),

[Continued on next page]

(54) Title: DYNAMIC ROUTE SELECTION FOR LABEL SWITCHED PATHS IN COMMUNICATION NETWORKS



(57) **Abstract:** The invention teaches a method for dynamically routing data in a Multi Protocol Label Switching network (Figure 3). A network of nodes operative to receive and transmit data are grouped into a plurality of clusters (152, 154, 156) which are hierarchically ranked (160); a leader node is also selected for each cluster. Data received at an ingress node is transmitted either to an egress node within the cluster, or, via the lead node, to another cluster of a different hierarchical rank. In some embodiments, a leader node contains in an associated routing table path information for each node in lower ranked clusters, as well as routing information for a leader node in at least one higher ranked cluster. The plurality of clusters may be organized hierarchically according to a rank determined by one or more Quality of Service QoS parameters, which may include jitter, loss, delay, and available bandwidth. The clusters may evolve along with traffic changes in the network; in particular, a given cluster may split or merge in response to changes in the applicable QoS metrics.

WO 2003/058868 A3

Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
25 March 2004

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US03/00163

A. CLASSIFICATION OF SUBJECT MATTER
IPC(7) : H04L 12/66, 12/28
US CL : 370/230.1,255,256,230.1,352,353,354,355,356,395.52,466
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 370/230.1,255,256,230.1,352,353,354,355,356,395.52,466

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2001/0019554 A1 (NOMURA et al) 06 September 2001 (06.09.2001), (Figures 4, 6, 8 and 9) and Page 2 [0022], [0027], [0052], [0055], [0064], [0075], [0077], [0085], [0087], [0092], [0096], and [0152].	1-60
Y	US 6,272,131 B1 (OFEK) 07 August 2001 (07.08.2001), column 14, lines 50-61, column 10, lines 56-63, and column 14, lines 45-60.	1-60

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
23 January 2004 (23.01.2004)

Date of mailing of the international search report
30 JAN 2004

Name and mailing address of the ISA/US
Mail Stop PCT, Attn: ISA/US
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450
Facsimile No. (703) 305-3230

Authorized officer
Kwang B. Yao
Telephone No. 703-305-3900

Form PCT/ISA/210 (second sheet) (July 1998)